Development of Quantum Mechanically Based Surface Models for Biological Fingerprinting

Dissertation
zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt
der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich
von
Anne Dara Bowen
aus den Vereinigten Staaten

Promotionskomitee:
Prof. Dr. Kim K. Baldridge(Vorsitz)
Prof. Dr. Jay S. Siegel
Prof. Andreas Wagner

Zürich, 2011

Abstract of the Dissertation

Development of Quantum Mechanically Based Surface Models for Biological Fingerprinting

by

Anne Dara Bowen

University of Zurich, 2011

Prof. Dr. Kim K. Baldridge, Chair

Most drug design efforts are hinged on the principle that structurally similar compounds tend to have similar physiochemical as well as biological properties. While this premise has lead to the development of many effective drugs, the mechanism by which bioactivity is induced by a drug for it's intended targets is often unknown. For computational studies, this leads to the question of how to define structural similarity algorithmically and validate the usefulness of the underlying assumptions. Drug design applications more and more require geometric data management along with physiochemical data handling and the 3D structures and surfaces of molecules become basic objects in molecular databases.

This thesis presents a computational methodology to more effectively deal with the large and complex chemical space governing CNS drug-receptor interactions. A particularly interesting and pertinent application example; the 5HT2a and it's receptor system and it's implications in affective disorders such as schizophrenia; has been chosen to exemplify the biological relevance of this work. This presentation begins with the historical development for what is known today about the 5HT2a as well as the computational approaches used to investigate this system and design selective compounds to understand the mechanism behind therapeutic drug action. Given the complexities of these types of receptor systems, rather than trying to assess the specifics of particular ligand-receptor interactions; the goal in this work is to assert a common reference from which it can be assumed that all things are equal and gauge the importance of interactions in a relative sense.

The computational strategies employed in this thesis to address the challenges of modeling these CNS drug-receptor interactions explores the correlation of activity with localized, global and fingerprint representations of molecular structure. The methodology presented offers 3 improvements to the computational infrastructure for the correlation

of structural data with CNS bioactivity: 1) extraction and communication of structure-activity relationship information that is encoded a physically significant QM based model 2) rigorous analysis of molecular shape and surfaces and the geometric data structures used to represent these ideas computationally 3) statistical methodology for correlation of molecular structure for bioactivity of these complex systems.

Zusammenfassung

Development of Quantum Mechanically based Surface Models for Biological Fingerprinting

von

Anne Dara Bowen

Universität Zürich, 2011

Prof. Dr. Kim K. Baldridge, Chair

Die meisten "Drug Designs" basieren auf der Idee, das chemisch strukturell hnliche Verbindungen, hnliche physiochemisches sowie biologische Eigenschaften haben. Diese Strategie hat zur Entwicklung vieler wirkungsvollen pharmaaktiven Substanzen gefhrt, jedoch ohne deren biologischen Wirkungmechanismus genauer zu verstehen. Im Bezug auf computergesttzte Methoden zur Bestimmung von Drogen-Rezeptoren Wechelswirkungen kommt die Frage auf, wie man strukturelle hnlichkeit algorithmisch definiert und die Verwendungsfhigkeit der zugrunde liegenden Annahmen validiert. "Drug Design" erfordern immer mehr geometrische Datenverwaltung zusammen mit der physiochemischen Datenbehandlung. 3D Strukturen der chemischen Verbindung und die Oberflchen der Molekle sind grundlegende Elemente in solch einer molekularen Datenbanken. Diese Doktorarbeit beschreibt eine computergesttzte Methode um effektiver den grosszgigien und komplexen Raum bei einem ZNS-Drogen-Rezeptor System zu modellieren um besser die ZNS-Drogen-Rezeptor Interaktion modellieren und verstehen zu knnen. Um die entwickelte Methode zu prfen wurde ein besonders interessantes und passendes Anwendungsbeispiel; das 5HT2a Rezeptor System als Prfsystem gewhlt. Die pathologischer Bedeutung des 5HT2a Rezeptor ist eine Fehlsteuerung biochemischer Vorgnge an den Serotonin-Rezeptoren die unteranderem zu Schizophrenie fhrt. Mit der anwendung der entwickelten computergesttzten Methode auf den 5HT2a Rezeptor, soll die biologische Bedeutung dieser Arbeit illustriert werden. Das Beispiel des 5HT2a Rezeptors wird durch einen historischen berblick eingeleitet, vorhanden Computerdatenstze werden disuktuiert und auf Verwendbarkeit der entwickelten computergesttzten Methode evaluiert. Zudem sollen Liganden modelliert werden, um die Ligand-Rezeptor System zu studieren. Bei einem solch grossen System, wie dem des 5HT2a Rezeptors ist der Mechanismus der Drogen-Rezeptor Interaktion sehr komplex. Das Ziel dieser Arbeit ist es eine allgemeine Referenz fr die rumliche Inter-

aktion zwischen Ligand-Rezeptor zu beschreiben. Die Computerstrategien, die in dieser Doktorarbeit eingesetzt wurden, um die Herausforderungen des Modellierens dieser ZNS-Drogen-Rezeptor Interaktionen zu adressieren, erforschen die Korrelation der Aktivitt; lokal und globalen und zieht den Fingerabdruckdarstellungen der molekularen Struktur mitein. Die prsentierte Methode zeigt drei Verbesserungenpunkte fr die Berechnung der Interaktion der strukturellen Daten mit der ZNS-Bioaktivitt dar: 1.) Extraktion und Kommunikation der Struktur-Aktivits-Verhltnisses, basierend auf einer quantenmechanischen Analyse 2.) Rigorose Analyse der Molekülstruktur im Bezug auf die geometrische Struktur, Oberflche. 3.) Statistische Methode zur Korrelation der Molekularen Struktur und der Bioaktivitt komplexer Drogen-Rezeptor Systeme.

Acknowledgement

The years spent working on my PhD changed me more than I ever would have anticipated at the outset. I have not only grown immensely as a person and as scientist, but my entire world view has expanded. This was largely due to my interactions with the people I worked with along the way, and I would like to take a moment to thank them. I feel so fortunate to have had so many wonderful people around my during the years I studied at the University of Zurich. I feel like these people contributed far more than just project guidance but helped me learn to be the best scientist I could be.

First, I would like to thank my advisor, Kim Baldridge. Without her, I would not even have begun. She gave me the courage to believe in myself and try something completely different. I have never looked back! From the moment I met her, her enthusiasm and love of science and mathematics inspired me to want to do more with my life than just work a regular job. I wanted to be like her, and find and pursue my passion. She supported my interests and patiently helped me through a thousand dead-ends in order to study the 5HT2a receptor system I have always been fascinated by. She has helped me think through scientific problems and turn many a vague notions into concrete sets of data. I really appreciate her kind but firm manner of dealing with all manner of problems (technical and personal). Through many tough spots, she helped me to do the best I could do and therefore be the best I could be. Even though I failed many times, she never made me feel like a failure. From her, I've learned how to stay calm, brush myself off, and try again tomorrow. I have never known a more genuinely caring individual and feel so fortunate to have had Kim as my advisor. What an adventure it was!

My father, Ed Bowen, for patiently and thoroughly explaining all statistical methods in this thesis to me. My Dad's excitement for statistics opened up a whole new world to me and inspired me to want to more. Thank you making the subject so much more than just number crunching! Thanks to both my parents (to my Mom, Achara, too) for always encouraging me to reach a little higher, even if it meant having me move far away for a lonnng time.

Sasha Buzko, for letting me use and contribute to the Sirius code base. I learned a lot from Sasha about good programming practices and appreciate that he was always willing to help me with any issue, from coding to docking.

My lab brothers and sisters: Becky, Yohann, Laura B., Laura Z. and Heidi, thanks for being such wonderfully easy-going and good-natured people. Laura B.'s enthusiasm for doing things well made everything more fun. I will always remember our MO problem sessions where she would bake some goodie to make it more enjoyable for us all. I can not imagine breaks without Becky's sarcastic humor and exciting stories.

My dearest friend and colleague, Fitore Kasumaj, many days I felt that she was the only soul in the universe who understood me. I am so thankful that I had a fellow PhD student who I could relate to. Even when everything seemed to be going wrong, we managed to find a way

to make each other feel better. Likewise, Celine Amoreira, always inspired me with her sense of determination and her hard work. Her charming smile made every day brighter! Girls, without our "Women in Science" meetings where would we be?

A very special thank you to Mike Packard, for believing in me and not letting me give-up. He always knew the right thing to make me forget about my worries and lose myself in laughter. I feel so fortunate that I had such a great buddy by my side through all this.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Motivation

The quest towards understanding the varied and complex mechanisms of ligand receptor interactions is greatly enabled by the inclusion of computational methodologies. Computationally derived data and models have the potential to bridge the gap between the actual experimental data and the detailed mechanism driving the chemistry of a system. Effectively bridging this gap is dependent on the appropriate choice of statistical methods to describe and interpret the correlation between the experimental and computational data as well as the quality of the data itself. The work described in this thesis is primarily concerned with the development of computational tools for the application of quantum mechanically derived descriptors to various chemical problems. This effort includes: 1) tools for the extraction and calculation of QM descriptors 2) methods for visualization and analysis of the QM data and finally 3) the infrastructure to apply these tools to the statistical analysis of biological and chemical data.

Quantitative structure activity relationship (QSAR) models are statistical models which correlate physiological properties or activities of molecules with some representation of their physical structure. To formulate a QSAR model, a set of variables, called 'descriptors', are chosen to characterize the molecular structure. These descriptors are used to develop a statistical relationship between the structure and properties or activities. The quality of the resulting model is highly dependent on the level at which the 'descriptors' are calculated. Descriptors for a molecule can be as crude as counting the number of hydrogens or as complex as a quantum mechanically calculated polarizability tensor. The quality of the resulting model will directly reflect the underlying representation. There exists an extremely large number of descriptors that can be computed for a given structure; however, not all are physically meaningful. The goal is to find fundamental physical properties or features of the set of molecules in question that can be used to describe differences in the observable of interest (e.g., affinity, activity, etc.). To be statistically manageable, typically many degrees of freedom must be removed. Traditional QSAR approaches often do this indiscriminately, and have been criticized for providing statistical models that do not express the underlying physicality of the mechanism. In this work, the use of quantum mechanically based descriptors provides a solid theoretical framework, but care still must be taken to ensure that these descriptors are modeled in a way that respects the

complex nature of the chemical or biological system. The quantum mechanical theory used for the descriptor calculation is presented in Chapter 2.

While it may be sensible to apply linear methods for prediction of physicochemical properties, conceptually there are issues associated with applying linear regression to relationships between structure and biological activity. It is not necessarily the case that a single linear model exists. This is particularly true in the presence of multiple binding modes, for example. If multiple binding modes are encompassed in a single linear model then a predictive outcome may only be due to binding contributions that stem from similar molecular features in both binding modes. Since the aim of this thesis is not to develop a predictive model of structure and activity per se, but rather to use statistical tools in an exploratory role, this consideration is less of a concern. However, nonlinear statistical methods are also explored in cases where linear models do not perform well, or where results are not easily interpretable. Specifically, a neural net infrastructure was developed and tested for molecule encoding, classification, and feature extraction. These results are compared with traditional factor analysis, as well as variable reduction tools such as principal component analysis, logistic regression, and partial least squares. neural nets are often criticized as analysis tools due to the difficulty in interpreting the results. One of the guiding goals of the work described in this thesis is the development of a visual analysis tool to aid in the interpretation of data from neural networks. The various statistical methods used in this thesis are described in Chapter 3.

When applied towards the study of ligand-receptor interactions, QSAR strategies are typically approached from either consideration of only the ligands, ligand-ligand approach, or modeling interactions between a ligand and receptor macromolecule, ligand-receptor approach. Ligand-receptor methods typically use the 3D coordinates of both the receptor and the ligand to directly analyze the interactions in the binding site. Molecular docking methods, for example, aim to calculate the optimized geometry for both the ligand and the receptor. In the absence of such three dimensional information of the receptor, a hypothetical receptor model can be deduced by analyzing series of molecules that interact with the receptor. Ligand-ligand approaches attempt to infer what the surface of a receptor might "look" like through such comparisons with known ligand targets. For example, a series of molecules that have high affinity for a receptor could be compared for similar structural features and electronic properties. This comparison can be carried out in a very reductionist manner by either trying to determine an analytical expression that represents some combination of features of the molecule that activate a receptor; or in a very complex way that attempts to model the receptor itself. The 'pharmacophore' provides a reduced geometric representation of the receptor, which includes only key molecular features found to be important for binding. For example, a pharmacophore for 5HT2a antagonist binding is shown at the bottom right of Figure 1.2. This pharmacophore consists of a nitrogen, two aromatic groups (Ar), and 3 distances. In contrast, the hypothetical receptor surface on the bottom right represents the entire 3D surface of the molecule. Many methods aim to compro-

mise between these two extremes. The motivation behind the methodologies developed in this thesis is to develop as accurate a model as possible without over-interpreting the data. While this may result in a less quantitative model in the predictive sense, such a technique should yield important insights towards development of more qualitative models where the underlying physicality is revealed.

Figure 1.1: Molecular Docking Example

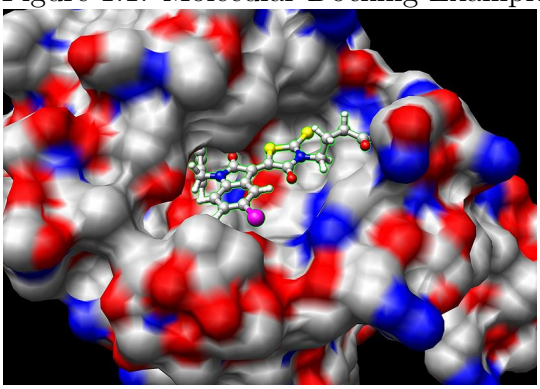
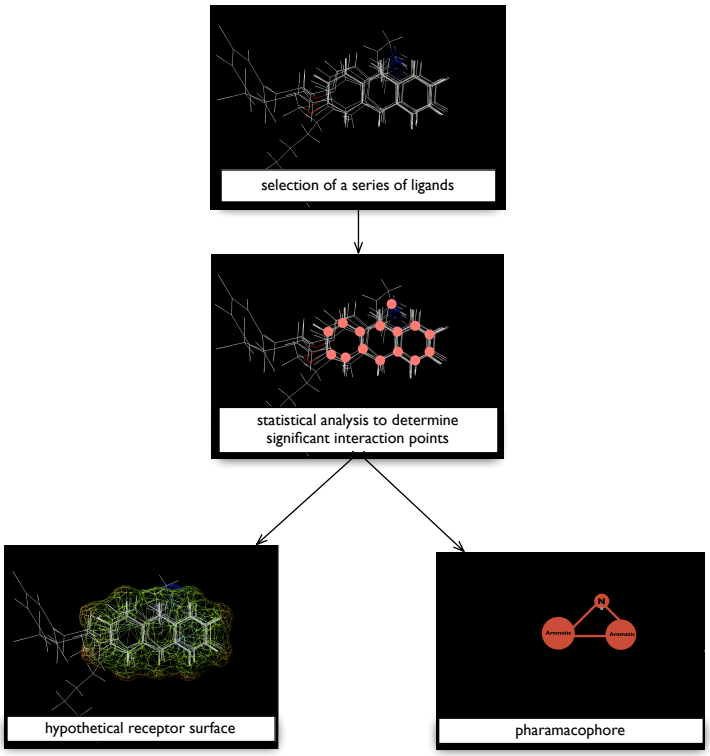


Figure 1.2: Summary of Ligand-ligand methods



Visualization provides a useful way to represent relevant structure and properties at the molecular surface. Rather than just displaying the computed data, the aim for the tools developed for this thesis was also to provide a computational basis for further analysis and model development. To accomplish this, the data structure that represents the visual surface was constructed in such a way that the underlying architecture is amenable to the algorithms and methods used in computational geometry. The Voronoi diagram and its geometric dual the Delaunay triangulation are mathematically well defined geometric structures that lend themselves naturally to the superposition of molecular properties required for the development of a hypothesis for a receptor surface model. Incidentally, the Voronoi diagram can be approximately computed using a type of unsupervised Neural Net, the Self Organizing Map (SOM). This connection between the Voronoi data structure and the SOM algorithm could be extended to provide a more intuitive interface for interpretation of the NN results. The utility of a mapping 3D data to a 2D SOM are explored to determine if they function as visual classification tools as well. This computational infrastructure and methodology is described in Chapter 8.

The application examples presented in this thesis were selected to address the issues in the application of QM derived data to statistical models. In Chapter 4, a rigorous analysis of the interactions between a series of molecules with an artificial cyclophane host directly examines factors important for ligand-receptor binding and the relevance of using QM descriptors. Chapter 5 further explores the utility of QM descriptors in representing reactivity patterns in a set of mono-substituted benzenes. Chapter 6 exemplifies the utility of the statistical approach used in a factor analysis of molecular strain in a series of triptycene dimers.

The task of trying to correlate molecular structure and properties with a specific experimental endpoint remains a daunting task, even after taking all the steps possible to reduce the experimental unknowns and chemical space to search. This task becomes even more challenging when applied to the study of ligand-receptor interactions in biological systems. G-Protein Coupled Receptors (GPCRs) are a large family of receptors activated by a broad variety of natural ligands. GPCRs are attractive targets for drug design because they act as receptors and signal transmitters, which allow a cell to communicate with the outside. This makes them ideal targets for therapeutic modulation of cellular responses. GPCRs are subclassified into families according to their pharmacological nature and sequence similarities. The serotonin receptor system represents one subfamily where structurally similar ligands demonstrate very different selectivities for the various subtypes of receptors. Moreover, a single ligand might activate several receptors. As is the case for almost all of the GPCR receptors, there is no experimental 3D structure available for any of the serotonin receptor subtypes. These receptors are major targets for pharmaceutical development, but the diversity, complexity and lack of an experimentally determined 3D structure has made the task of developing suitably selective agonists and antagonists extremely challenging. Chapter 7 presents the results of applying computational tools developed in this thesis to a series of selective antagonists of the 5HT2a subtype of the serotonin

receptor family, in attempt to break through some of these challenges and offer mechanistic insight to the problem.

Chapter 2

QM Analysis of Molecular Interactions

Quantum chemically derived descriptors can, in principle, express all of the electronic and geometric properties of molecules and their interactions. QM descriptors can characterize the properties of the entire molecule (global descriptors) or they can be also partitioned on the basis of atoms or groups (local descriptors), allowing the description of various molecular regions separately. QM properties can also be used to calculate a "fingerprint" representation which reduces the information content of the entire molecule to some characteristic mapping of features. This thesis investigates the use of gas phase local and global QM descriptors derived from the molecular electron density, as well as solvent screening charge density taken as a fingerprint descriptor.

The computation of the QM descriptors described in this chapter required the development of custom software to parse the output from the *ab initio* electronic structure package GAMESS[1] to extract, for example, molecular orbital eigenvalues, eigenvectors and other properties, for further computation. The data structures, computational methods, and infrastructure, are described separately (Chapter 8) so as to not disrupt the flow of the theoretical background presentation here.

2.1 Electronic Structure Theory Calculation

All of the information regarding the structure, properties and energetics of a molecule can be determined from its wavefunction Ψ, as determined from solving the Schroedinger's equation $E\Psi = \hat{H}\Psi$. $\hat{H}$ is the Hamiltonian operator and represents the energy of the nuclei and electrons, Ψ is the representation of the molecule, and E is the resulting energy of the molecule. There are several approximations that must be made to make solving this equation computationally feasible. The starting point for most quantum chemistry calculations is to represent the wavefunction, psi, describing the molecule, as a Linear Combination of Atomic Orbitals (LCAO). A Molecular Orbital (MO) is a composite of weighted Atomic Orbitals (AOs) which collectively

Figure 2.1: MO Eigenvectors

$$\begin{bmatrix} & \mathbf{MO_1} & \mathbf{MO_2} & \mathbf{MO_3} & ... & \mathbf{MO_N} \\ \mathbf{AO_1} & c_{11} & c_{12} & c_{13} & ... & c_{1N} \\ \mathbf{AO_2} & c_{21} & c_{22} & c_{23} & ... & c_{2N} \\ \mathbf{AO_3} & c_{31} & c_{32} & c_{33} & ... & c_{3N} \\ ... & ... & ... & ... & ... & \\ \mathbf{AO_N} & c_{N1} & c_{N2} & c_{N3} & ... & c_{NN} \end{bmatrix} \tag{2.1}$$

define the shape and spatial density of the electrons in a molecule. The outcome of a quantum chemistry calculation based on the Hartree-Fock/Self Consistent Field method [2] includes a matrix of eigenvectors and eigenvalues corresponding to the molecular orbitals of the molecule(Figure 2.3). The eigenvectors are a square matrix where the dimensionality is equal to the total number of atomic orbitals. The rows correspond to each of the repsentative atomic orbitals in the linear combination of atomic orbitals, and the columns to the molecular orbitals of the molecule. For example, Equation (2.2) shows the coefficients from the first column in Figure 2.3, representing the first molecular orbital. Ψ represents the spatial form of the molecular orbital, and ϕ is a function that describes a particular atomic orbital.

$$\Psi(1) = c_{11}\phi(1) + c_{21}\phi(2) + c_{31}\phi(3) + ... + c_{N1}\phi(N) \tag{2.2}$$

The eigenvalues are represented as a diagonal matrix (Figure 2.4), where each element on the diagonal line (e_{ii}) is the energy of the corresponding orbital (column) i in the eigenvector matrix. The columns of the eigenvectors table are always ordered by the corresponding eigenvalues, so that the first MO is the lowest in energy. These energy levels and electron occupancies of the molecular orbitals are frequently plotted to gain chemical insight on the reactivity of the system, as shown with benzene as an example in Figure 2.3. In order to plot the shape of the molecular orbitals, it is necessary to know the mathematical expression which describes its intensity at all points in 3-dimensions in addition to the wavefunction. Slater functions (Equation (2.3)) are the most conceptually simple manner in which the individual atomic orbitals can described. The Cartesian variables (x, y, z) are the displacements from the center of the atomic orbital (the center being the position of the atom) and r is the magnitude of distance from the center. A, b, c and zeta are parameters for the orbital, which must be optimized for each atom type and basis set, and N is a normalization constant. The Cartesian terms arise from a, b and c when they are nonzero, which represent the angular momentum along the particular axis/axes. For example, p-orbitals are described where just one of a, b or c is equal to 1, depending on whether it is px, py or pz. *Ab initio* methods typically do not use Slater functions because it is necessary to evaluate a large number of 3D spatial integrals during the calculation. In order to make the integral calculations feasible, a series of contracted Gaussian-type functions (where

Figure 2.2: MO Eigenvalues

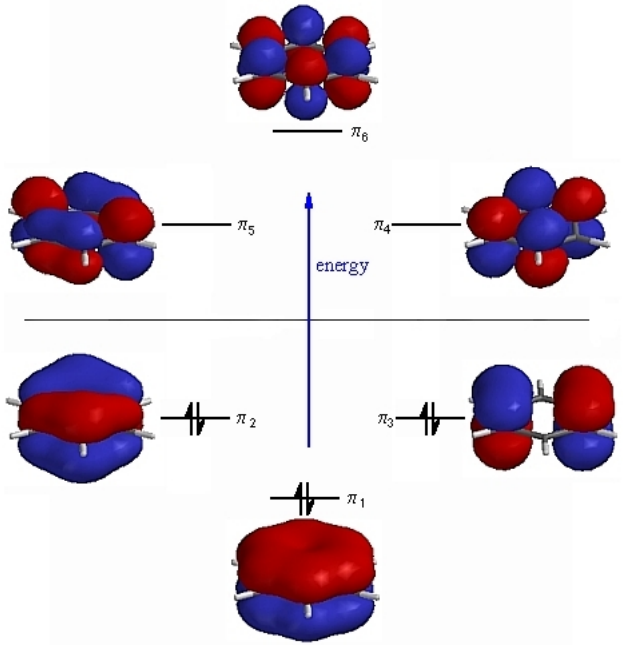
$$\begin{bmatrix} & \mathbf{1} & \mathbf{2} & \mathbf{3} & ... & \mathbf{N} \\ \mathbf{1} & e_{11} & & & & \\ \mathbf{2} & & e_{22} & & & \\ \mathbf{3} & & & e_{33} & & \\ ... & & & & & \\ \mathbf{N} & & & & e_{NN} & \end{bmatrix} \tag{2.4}$$

the exponential term is e^{-kr^2} are used instead.

$$\phi(i) = Nx^a y^b z^c e^{-\xi r} \tag{2.3}$$

Ordering the MOs by energy makes it straightforward to populate the electron occupancy of the molecular orbitals: if there are M electrons in a closed-shell system, then the first M/2 molecular orbitals are occupied, with two electrons each. The last of these to be filled is referred to as the Highest Occupied Molecular Orbital (HOMO). The molecular orbitals that are not filled are referred to as "virtual", or empty, orbitals. The first virtual orbital is referred to as the Lowest Unoccupied Molecular Orbital (LUMO). In 1952, Kenichi Fukui[3] realized that a good approximation for reactivity could be found by looking at only these HOMO/LUMO "frontier" orbitals. This application of MO theory to describing HOMO / LUMO interactions is known as Frontier Molecular Orbital theory.

Figure 2.3: Molecular Orbitals Eigenvector and Eigenvalue



Chemical interactions are typically thought of as being either electrostatic (polar) or orbital

(covalent). The electrical charges in the molecule are primarily responsible for electrostatic interaction, as such, charge based descriptors, such as the molecular electrostatic potential, are frequently used as indicators of weak intermolecular interactions. The covalent donor-acceptor interactions are thought to be characterized by the descriptors that are related to the frontier orbitals, such as the frontier orbital density and the superdelocalizability.

2.2 Molecular Orbital based Descriptors

The HOMO is the highest molecular orbital energy level that contains electrons. When a molecule acts as an electron-pair donor in bond formation, the electrons are supplied from the molecule's HOMO. How readily this occurs is reflected in the energy of the HOMO. Molecules with a high HOMO are more able to donate their electrons and are hence relatively reactive compared to molecules with low lying HOMO, making the HOMO energy a measure of the nucleophilicity of the molecule. The HOMO energy is directly related to the ionization potential and characterizes how susceptible the molecule is toward an attack by an electrophile.

The LUMO is the lowest energy level in the molecule that contains no electrons. When a molecule acts as an electron-pair acceptor the incoming electron pairs are received into its LUMO. Molecules with low-lying LUMOs are more able to accept electrons than those with high LUMOs; the LUMO is a measure of the electrophilicity of a molecule. The energy of the LUMO is directly related to the electron affinity and characterizes how susceptible the molecule is to attack by a nucleophile.

The difference between the HOMO and the LUMO eigenvalues (E_HOMO E_LUMO) is also an important stability index. This quantity is called the HOMO-LUMO gap. A large HOMO-LUMO gap implies high stability (lower reactivity). The HOMO-LUMO gap is directly related to the chemical concepts of hardness and softness when describing chemical species. Hard molecules have a large HOMO-LUMO gap while soft molecules have a small HOMO-LUMO gap. 'Hard' applies to species that are small, have high charge states, and are weakly polarizable. 'Soft' applies to species that are large, have low charge states, and are strongly polarizable.

Given the importance of the HOMO and LUMO gap in governing chemical reactivity, an important reactivity descriptor would involve these two orbitals. In fact, as shown by Fukui and in many subsequent studies, the difference in energy between the HOMO and the LUMO, the HOMO-LUMO gap, has been shown to be an important index for stability, or low chemical reactivity. A large HOMO-LUMO gap is indicative of high stability, while a small HOMO-LUMO gap is indicative of low stability, or high chemical reactivity. The symmetry and charge distribution within these orbitals are also important factors for structure and reactivity. Another relevant set of reactivity descriptors are the concepts of hardness and softness of the frontier orbitals.[4] In this paradigm, the index of chemical stability and reactivity is related to the global hardness of the electronic structure. A larger degree of hardness implies more stability, less

ability of the electronic structure to move or be polarized. For example, as a system moves away from equilibrium (as in a chemical reaction), the hardness value would decrease, and stability would decrease. Softness is the reverse of hardness, and implies that the electron density is more susceptible to being polarized. In other words an increase in softness of a molecule implies and increase in chemical reactivity. The global hardness (Equation (2.5)) and softness (Equation (2.6)) are computed using Koopmans Theorem.[5] The hardness of the molecule is the reciprocal of the respective softness (S) of the molecule, determined by Equations (2.5) and (2.6) respectively.

Another relevant set of reactivity descriptors are the concepts of hardness and softness of the frontier orbitals.[4] In this paradigm, the index of chemical stability and reactivity is related to the global hardness of the electronic structure. A larger degree of hardness implies more stability, less ability of the electronic structure to move or be polarized. For example, as a system moves away from equilibrium (as in a chemical reaction), the hardness value would decrease, and stability would decrease. Softness is the reverse of hardness, and implies that the electron density is more susceptible to being polarized. In other words an increase in softness of a molecule implies and increase in chemical reactivity. The global hardness (Equation (2.5)) and softness (Equation (2.6)) are computed using Koopmans Theorem.[5] The hardness of the molecule is the reciprocal of the respective softness (S) of the molecule, determined by Equations (2.5) and (2.6) respectively.

$$\eta(hardness) = -1/2(\varepsilon_{HOMO} - \varepsilon_{LUMO}) \tag{2.5}$$

$$\eta = \frac{1}{2S} \tag{2.6}$$

2.2.1 Frontier Orbital Densities (FOD)

According to frontier orbital theory,[3] the majority of chemical reactions take place at the position and in the orientation where the overlap of the HOMO and LUMO can reach a maximum. For electron donors, the HOMO density is critical to the charge transfer and this quantity is termed the electrophilic electron density. For electron acceptors the LUMO density is important and is termed the nucleophilic electron density.

$$f_n^E = \sum_{0}^{nOcc-1} (C_{HOMO,n})^2 \tag{2.7}$$

$$f_n^N = \sum_{nOcc}^{nBasis} (C_{LUMO,n})^2 \tag{2.8}$$

In the equations above $C_{HOMO,n}$ and $C_{LUMO,n}$ are the coefficients of the atomic orbital χ_n in the HOMO and LUMO respectively. These frontier electron densities can only be used to describe the reactivity of different atoms on the same molecule, however. In order to compare

reactivities on different molecules, the frontier electron density value must be normalized by dividing by the corresponding energy eigenvalue of the respective molecular orbital. These normalized FODs are also termed approximate superdelocalizabilities.

2.2.2 Approximate Superdelocalizability

As was mentioned in the preceding section, the frontier densities are only useful for comparisons of atoms in a single molecule. In order to make comparisons across molecules, the respective HOMO or LUMO sums must be divided by the respective energy eigenvalue. The normalized frontier orbital densities are also called the approximate superdelocalizabilites. Superdelocalizability and the relation to approximate superdelocalizability will be discussed properly in the next section.

The assumption for the nucleophilic and electrophilic frontier orbital densities is that the greatest interaction will occur at the site of largest orbital density (c^2 is largest). Since the energy eigenvalue of the HOMO is a negative value, the sign of the electrophillic superdelocalizability will always be negative. A more negative value indicates that the atom at that position is more susceptible to electrophilic attack. Notice that the equation for nucleophilic superdelocalizability has a negative sign so it also will be a negative quantity. A more negative value also indicates greater susceptibility to nucleophilic attack. It is easiest just to think of the overall magnitude of the descriptor as being of importance. Another descriptor that is sometimes used is the sum of the approximate superdelocalizabilities for a molecule or a molecule fragment, so these were calculated as well for evaluation.

$$F_n^E = \frac{f_n^E}{\varepsilon_{HOMO}} \tag{2.9}$$

$$F_n^N = -\frac{f_n^N}{\varepsilon_{LUMO}} \tag{2.10}$$

2.2.3 Superdelocalizability

The superdelocalizability is conceptually similar to the normalized frontier orbital density and can also be used to characterize atoms on different molecules. The quantity involves all of the respective MO eigenvectors rather than just the HOMO or the LUMO. For example, in the case of the electrophilic superdelocalizability, the expression involves the molecular orbitals from 1 to the number of occupied orbitals, and in the case of the nucleophilic superdelocalizability, all of the unoccupied orbitals are used. In principle, the use of such quantities enable one to avoid the problem where the density is not just in the HOMO but distributed over several of the lower lying levels.

Similar to the other MO based descriptors described thus far, the superdelocalizability concept is based on the idea proposed by Fukui in 1957[3] that the interaction of the molecular

orbitals of two reactants are a mutual perturbation, with the relative energetics of the two orbitals changing together to maintain a similar degree of overlap as the reactants approach each other. The assumption is made that the greatest interaction will occur at the site of largest orbital density (c2 is largest). The ej term accounts for the delocalizability. For low-lying levels the energy will be large and negative. This can be interpreted as meaning the electrons are tightly held and not very delocalizable. For higher energy occupied states (such as the HOMO) ej is much smaller; the electrons in the higher-energy orbitals are less tightly bound and consequently more delocalizable. The higher energy levels will dominate the superdelocalizability term. If the electrons of the frontier orbital predominate in those interactions, the one-orbital analog of superdelocalizability, i.e. the superdelocalizability calculated based only on the frontier orbital, can be used. This is the same as the normalized frontier orbital density. Summing S for all atomic positions of a molecule gives a metric of electrophilicity, which may be used to predict relative reactivity in a series of molecules.

The superdelocalizability is conceptually similar to the normalized frontier orbital density and can also be used to characterize atoms on different molecules. It uses all of the respective MO eigenvectors rather than just the HOMO or the LUMO. So for the electrophilic superdelocalizability, the expression uses the molecular orbitals from 1 to the number of occupied orbitals and for the nucleophilic superdelocalizability, all of the unoccupied orbitals are used. In principle, an approach such as this should help to avoid the problem seen the benzeneNO2 electrophilic frontier orbital density; where the density was not just in the HOMO but distributed to some of the lower lying levels.

Similar to the other MO based descriptors described thus far, the superdelocalizability concept is based on the idea proposed by Fukui in 1957[3] that the interaction of the molecular orbitals of two reactants are a mutual perturbation, with the relative energetics of the two orbitals changing together to maintain a similar degree of overlap as the reactants approach each other. The assumption is made that the greatest interaction will occur at the site of largest orbital density (c2 is largest). The ej term accounts for the delocalizability. For low-lying levels the energy will be large and negative. This can be interpreted as meaning the electrons are tightly held and not very delocalizable. For higher energy occupied states (such as the HOMO) ej is much smaller; the electrons in the higher-energy orbitals are less tightly bound and consequently more delocalizable. The higher energy levels will dominate the superdelocalizability term. If the electrons of the frontier orbital predominate in those interactions, the one-orbital analog of superdelocalizability, i.e. the superdelocalizability calculated based only on the frontier orbital, can be used. This is the same as the normalized frontier orbital density. Summing S for all atomic positions of a molecule gives a metric of electrophilicity, which may be used to predict relative reactivity in a series of molecules.

$$S_{n,electrophilic} = 2 \sum_{j=0}^{nOcc-1} \frac{c_{jn}^2}{\varepsilon_j} \tag{2.11}$$

$$S_{n,nucleophilic} = - \sum_{j=nOcc}^{nBasis} \frac{c_{jn}^2}{\varepsilon_j} \tag{2.12}$$

$$\rho(r) = \sum_i \sum_j P_{i,j} \chi_i^*(r) \chi_j(r) \tag{2.13}$$

$$P_{i,j} = \sum_{a=0}^{nOcc-1} n_a C_{i,a}^* C_{j,a} \tag{2.14}$$

$$\rho(r) = \sum_{a=0}^{nOcc-1} n_a \Phi_i^*(r) \Phi_j(r) \tag{2.15}$$

$$\chi(r) = N x^a y^b z^c e^{-\zeta r} \tag{2.16}$$

S_n = superdelocalizability at position of atom

ε_j = bonding energy coefficient in jth MO (eigenvalue)

c_{ja} = molecular orbital coefficient at atom a in the HOMO (nucleophilic) or LUMO (electrophilic)

m = index of the HOMO (nucleophilic)

2.3 Electron Density Based Descriptors

The total density of the orbital at a point in space is the sum of those of the constituent atomic orbitals at that point, multiplied by the weighting coefficient taken from the eigenvector matrix. Each of the atomic orbitals is represented by a mathematical expression that describes its intensity at all points in 3-dimensions (Equation 2.18. For purposes of graphically plotting the shapes of the molecular orbitals, these functions must be known, in addition to the wavefunction.

$$\rho(r) = \sum_i \sum_j P_{i,j} \chi_i^*(r) \chi_j(r) \tag{2.17}$$

$$\chi(r) = N x^a y^b z^c e^{-\zeta r} \tag{2.18}$$

2.4 Charge Based Descriptors

2.4.1 Molecular Electrostatic Potential

The molecular electrostatic potential is highly informative in terms of the nuclear and electronic charge distribution of a molecule and correlates with dipole moment, electronegativity, and partial charges, and provides a way to understand the relative polarity of a molecule. The MEP ε_p is defined as the energy of interaction of a positive point charge Z_A located at a some point r_1 with the nuclei and electron of a molecule. P_{ij} are the elements of the density matrix and χ are the atomic basis functions. R_A are the nuclei positions (for every atom in the molecule). The electrostatic potential represents a balance between the repulsion of the point charge by the nuclei and the attraction of the point charge by the electrons.

$$\varepsilon_p = \sum_{A}^{nuclei} \frac{e^2 Z_A}{4\pi\varepsilon_0(r_1 - R_A)} - \sum_i \sum_j P_{ij} \int \frac{\chi_i^*(r)\chi_j(r)}{(r_1 - r)} dr \tag{2.19}$$

2.4.2 Partial Charge

Partial atomic charge can be representative of molecular polarity in terms of charge build-up or depletion on individual atoms. For example, the hydrogen bonding ability of water is always described in terms of the oxygen atom having a partial negative charge, while the hydrogen atoms have a partial positive charge. These quantities are determined by some method of partitioning of the electronic distribution to give the charge per atoms in the molecule. There are several different methods for partitioning that are described in detail in Chapter 9 of Cramer's Computational Chemistry book.[2] In the current work, CHELPG (CHarges from Electrostatic Potentials using a Grid based method) derived charges as computed in GAMESS are used. In this method of partitioning, the atomic charges are fit to reproduce the molecular electrostatic potential around the molecule.

2.4.3 Solvent Screening Charge Density

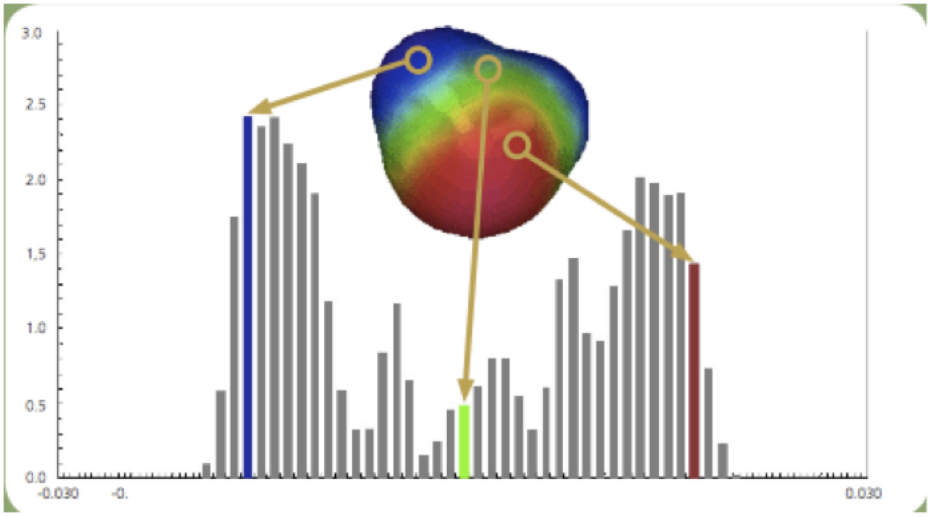
Continuum solvation models describe the electrostatic behavior of a solvent via a dielectric continuum. Continuum solvation methods model a molecule in solution by placing it inside a cavity formed within a homogenous medium, taken to be the solvent. Generation of such a cavity involves dividing the surface describing the boundary to the dielectric, into a grid of triangles. The screening charges on the cavity boundary resulting from polarization by the solvent are calculated iteratively using a self-consistent field method. The screening charge of each segment patch divided by the area of each path represents the screening charge density σ, which is used to construct the charge density fingerprint (units of chargenm$\hat{2}$). The fingerprint itself is a histogram where each bin represents the number of segment patches with the particular charge.

$$p(\sigma) = \frac{n_i(\sigma)}{n}$$

The numerator $n_i(\sigma)/$ is number of segments with surface-charge density σ and the denominator 'n' is the total number of segments.

In essence, the full 3D charge density information on the molecular surface is reduced to a histogram in this method, and is indicative of how much of a surface is polar and how much is nonpolar [σ-dσ/2, σ + dσ/2], and to what extent. In this way, the shapes of these histograms provide polarity profiles for molecules. For example, a polar molecule such as water has two peaks on either size of zero, a positive peak associated with the response of the negative charge on the oxygen, and a negative peak associated with the response of the partial positive charge of the hydrogens. A nonpolar molecule such as hexane would have just one peak centered around 0. In water, the charge density profile (2.4 shows approximately same amount of strongly negative and strongly positive surface area enables energetically very favorable pairings of positive and negative surfaces and formation of strong hbs without any lack of adequate partners. The scale in 2.4 is in Angstroms. The peaks at -1.5 nm in water correspond to the strongly positive H ions in water and the +1.5nm in water indicate the strongly negative polar regions of the electron lone pairs.

Figure 2.4: Example of Charge Density Histogram



Chapter 3

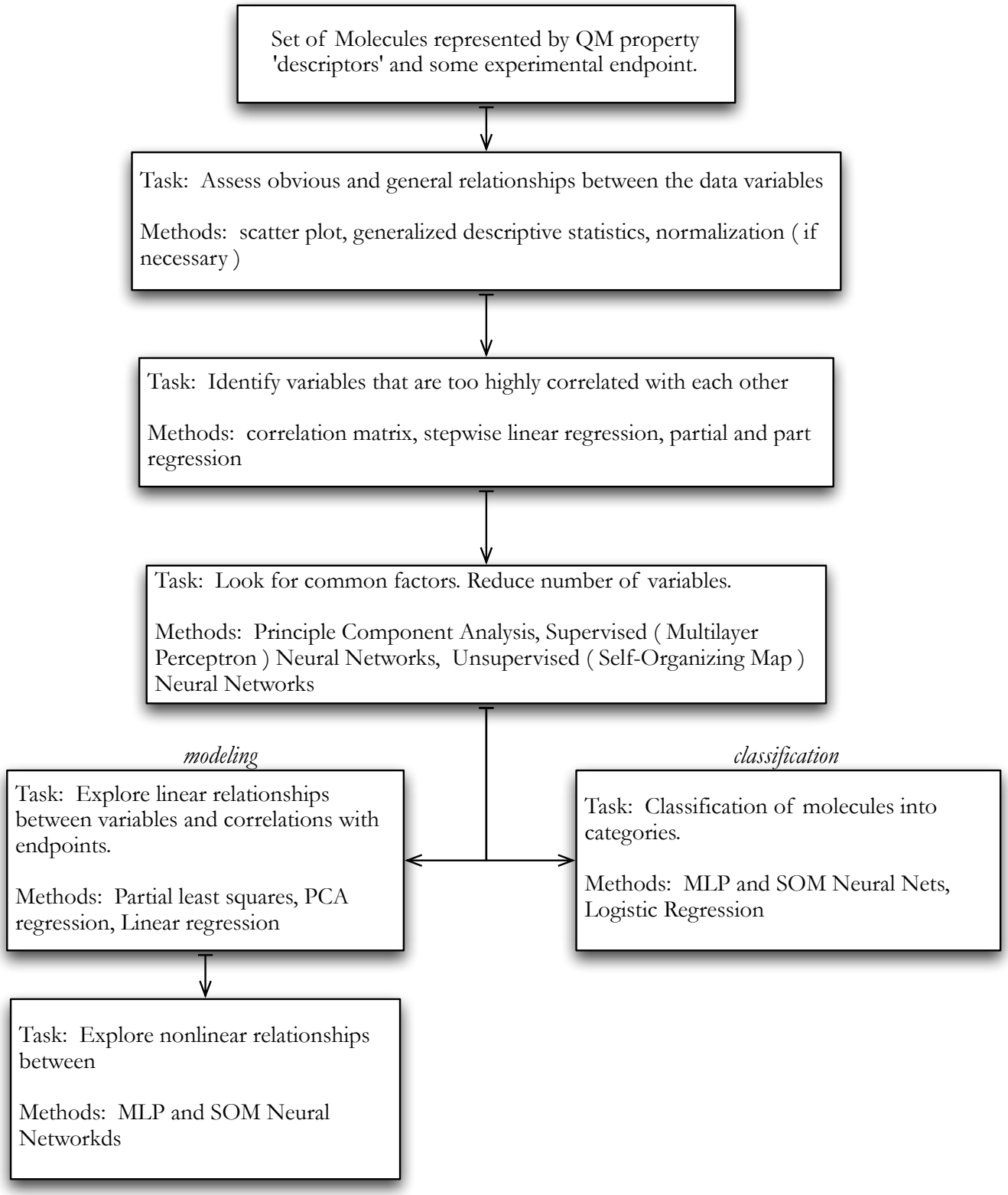
Statistical Methodologies for QSAR

The goal of a QSAR model is to extract information from a set of numerical descriptors that characterize a molecule's reactivity and use this to develop a quantitative relationship between structure and some endpoint. The previous chapter explored the importance of the using high-quality descriptors to represent the molecule and here the importance of using appropriate statistical methods will be explored.

Regression techniques assume a linear relationship between the biological activity and one or more descriptors. Regression analysis becomes problematic when there are a large number of variables. A frequent problem with large descriptor sets, is the redundancy in information when descriptors are correlated. For example, in the previous chapter all of the descriptors based on the frontier orbitals were highly correlated. Latent variable techniques, such as Principle Component Analysis and Partial Least Squares address this issue. A pervasive problem when using computational methods to generate descriptors is that there is frequently more descriptors than compounds. This introduces the possibility that correlations observed may be chance correlations.

The overall statistical methodology is outlined in Figure 3.1. The first step in the analysis is to explore the data to get a general idea of any patterns within it and determine if any sort of standardization or normalization scheme is required. After this, any variables that are too highly correlated with each other are identified and it is determined if they need to be removed from the study, similarly outliers are flagged for further consideration. Once any problematic cases are dealt with, common factors and patterns are identified. At this point, the methods used depend on if the final goal is to develop a predictive model or to explore the data further in an exploratory context. The statistical methods used for each of these step will be described in this chapter.

Figure 3.1: Overall Statistical Methodology



3.1 Covariance and Correlation

Variance is a measure of dispersion around the mean of a random variable. When a sample is used the sample variance ($s_{i,j}$) is given by the equation (3.1). The sample standard deviation is defined as the square root of the sample variance. Covariance measures the linear association between two or more random variables. A larger absolute value for the covariance implies a stronger linear relationship between the two variables. It is important to note, however, that the covariance does not necessarily capture nonlinear relationships.

$$cov(x, y) = s_x s_y = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1 s_x s_y} \tag{3.1}$$

One problem with the use of covariance to compare the linear relationships of more than one independent variable with a dependent variable is that the covariance is highly dependent on measurement scale. So, one independent variable having a larger covariance than another with the dependent variable does not imply that it has a stronger linear relationship. A solution to this problem is to use the sample correlation coefficient which , as shown in equation (3.1). , is simply the sample covariance divided by the product of the sample standard deviations. This puts the covariance values into a standard set of units (between -1 and +1) that is known as the Pearson product-moment correlation coefficient. A coefficient of +1 indicates the variables are perfectly positively correlated and a value of -1 indicates the variables are perfectly inversely correlated. The correlation matrix of n variables is an n x n matrix whose i,j entry represents the correlation between X_i and X_j. the diagonal of the correlation matrix will always be 1 since every variable is perfectly correlated with itself.

$$r = \frac{cov(x, y)}{s_x s_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1 s_x s_y} \tag{3.2}$$

The square of the correlation coefficient (r^2) measures the actual amount of variation held in common between the two variables and is useful in interpreting the correlations. The reliability of the correlation coefficient is assessed by the significance level (p), which is dependent on sample size and assumptions regarding the distribution of residuals. In general, the larger the sample size and the closer to the distribution of residuals is to the Normal probability distribution, the more reliable the correlation coefficient is as an indicator of linear association. A highly significant result is reflected in a low p value, which indicates that the result is unlikely to have resulted by random statistical variation.

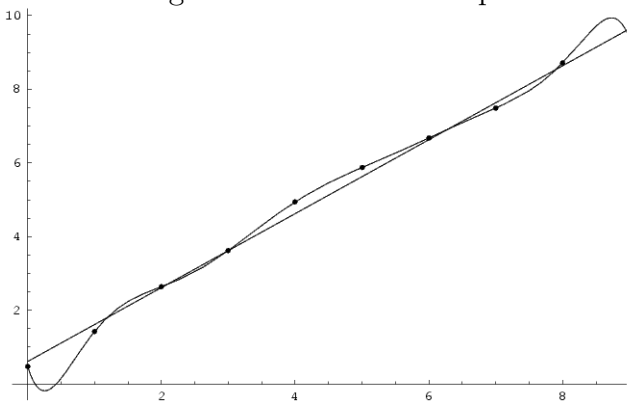
3.2 Part and Partial Correlations

In multiple linear regression analysis, where a large number of independent variables are often involved, partial and part correlations are often computed in addition to the multiple correlation coefficient previously described. Partial correlation is a measure of the correlation between two variables in which the effects of other variables are removed from both the other independent as well as the dependent variables. In effect; the influence of everything else in the model is subtracted out. Part (also called semi-partial) correlation removes the effect on an independent variable from the other independent variables, but does not remove the effect of the other independent variables on the dependent variable. Examining the part and partial correlations within the context of a linear regression analysis can often help illuminate complex dependencies among the variables[6]

3.3 Regression

Correlation analysis is a good first step in selection of candidate variables for a more predictive or explanatory model such as a Regression model, which involves fitting a predictive model to a data set and using that model to predict values of the outcome (dependent) variable from the predictor (independent) variables. In the simplest case, this "predictive model" could just be the equation for a line. If there are more variables than cases, care must be taken to ensure the model is not overfit.For example, it is possible to fit a line of order emph(n-1) from any emphn points. As can be seen in Figure 3.2 the polynomial function passes through each data point but would do a poorer job of extrapolating data if it were used as a regression curve. Also, a regression analysis will fail if the independent variables are too highly correlated with each other so care must be taken to ensure that perfectly correlated variables are removed before the analysis.

Figure 3.2: Overfit example



3.4 Principle Component Analysis

As was seen in the beginning of the chapter, the correlation matrix gives the correlations between all the variables of interest. Sometimes there are clusters of high correlation in the subsets of variables suggesting that those variables could be measuring some aspect of an underlying dimension. These underlying dimensions are called "factors" or "latent variables". By reducing the set of interrelated variables into a smaller set of uncorrelated factors, a factor analysis can explain the maximum amount of common variance using a reduced number of factors. There are many different methods for performing a factor analysis but they have the same basic principle: they transform the original variables into a smaller set of variables which retains much of the information contained in the original set. Both Factor Analysis and Principal Components Analysis (PCA) derive new variables from linear combinations of the original ones.

PCA determines the underlying factors by utilizing the eigenvectors and eigenvalues of the correlation matrix. First,the eigenvectors are ranked according to their eigenvalues. The components of the eigenvectors are used as the coefficients to form linear combinations of the original variables. The new variable (principal component) formed from the eigenvector having the the largest eigenvalue being referred to as the first principle component, and so forth. There are as many principal components as original variables if it is desired to compute them all. However, most of the variation in the data can often be explained with only a few principal components. Frequently, a calculation of what percentage of the total variance is represented by each eigenvector is performed to determine how many eigenvectors need to be considered to account for the bulk of the total variance. PCA is primarily a data reduction technique and it involves only the independent variables. Although the principle components can be used as new variables in a linear regression model,the strength of relationship of the principal components with the dependent variable is not reflected in the principle components, and such a model may not be very predictive. When prediction is desired the Partial Least Squares method is often much more appropriate.

3.5 Partial Least Squares

The goal of Partial Least Squares (PLS) is to predict the dependent variables Y from the independent variables X and describe their common structure. Similar to PCA, PLS can be used to identify underlying dimensions (factors) that can be used as a way to reduce the data set from a group of interrelated variables into a smaller set of uncorrelated factors. Often, it is possible to interpret these extracted factors in terms of the underlying physical system. PLS can be used as a variable reduction technique and be interpreted in a similar fashion to PCA. PLS differs from PCA because both the X variables (independent variables or factors) and the Y variables (dependent variables or responses) are reduced to principal components. The PCA algorithm is only concerned with maximizing variance in the X variables and does not consider

the response at all. To construct the principle components of X the PLS algorithm iteratively maximizes the strength of the relation between successive pairs of the X and Y component scores by maximizing the covariance of each X-score with the Y variables. The methodology employed in the PLS algorithm gives PLS some advantages over linear regression type methods: 1) the X components used to predict Y will be orthogonal so the method will not fail if the original X variables are multi-collinear and 2) for the Regression only a few of the components are used in the prediction so PLS will not overfit the data even when there are more variables than cases/observations.

3.6 Neural Networks

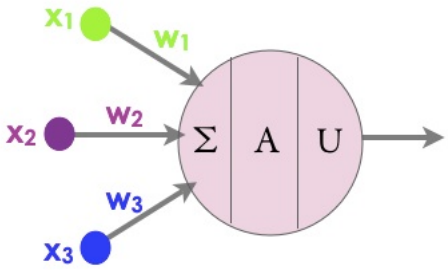
Artificial Neural Network (ANN) algorithms are designed to "learn" data in a way that emulates learning in the brain. An artificial neural network consists of a number of `neurons` that receive data, process the data, and output a signal. A `neuron` is essentially a regression equation with a non-linear output. When more than one of these neurons is used, non-linear models can be fitted. There are two basic types of ANN algorithms, supervised and unsupervised. Both supervised and unsupervised neural nets have the same basic neuron structure shown in Figure 3.3. The network receives a set of inputs (x1, x2 and x3) which are multiplied by each neuron's weights (w1, w2 and w3). These products are summed for each neuron by some summing function $\sum$ and a non-linear transfer function A called the activation function is applied. The transformed sums are then multiplied by the output weights where they are summed once more, transformed, and interpreted by the update function U.

In both cases, the neural networks are presented with a series of patterns during training. The supervised ANN "learns" by adjusting its connection weights to minimize error (predicted versus actual binding affinity for example) while the unsupervised ANN adjusts its connection weights until it settles into some stable state. The supervised ANN algorithms are typically used for modeling in a very analogous way to linear regression but do not require or assume a linear relationship. The advantage of using a NN is that they are able to model a wide set of functions without knowing the model a priori and they are typically used when the problem is not understood well enough to write a procedural program. The unsupervised network is typically used for variable reduction (mapping) and classification tasks.

3.6.1 Supervised NN: Multi-Layer Perceptron

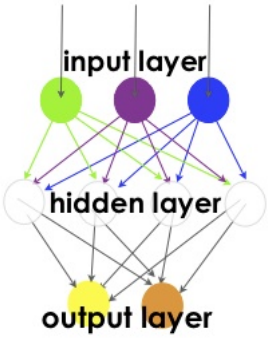
The multilayer perceptron (MLP) is a very common ANN architecture. An MLP consists of multiple layers of neurons connected unidirectionally. The basic structure of a MLP neural net is shown in Figure 3.5. The network receives a set of inputs which are multiplied by each neuron's weights. These products are summed for each neuron and a non-linear transfer function is applied. The transformed sums are then multiplied by the output weights where

Figure 3.3: Neuron



they are summed once more, transformed, and interpreted. The error (difference between the desired output and the network's predicted output) is calculated and is propagated backwards through the network, adjusting the weights so that the next time the network sees the same input pattern, it will come closer to the desired output. The patterns are presented over and over to the network until the error is within a certain range.

Figure 3.4: Multilayer Perceptron

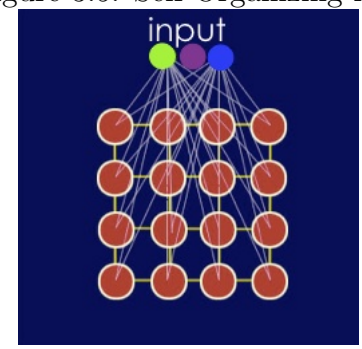


3.6.2 Unsupervised NN: Self Organizing Map

A self-organizing map (SOM) is a type of ANN that is trained using unsupervised learning. Typically, the result is a two-dimensional discretized "map" representation of the input space of the training samples. There are three phases of SOM operation: training, mapping and testing. Before training begins, the weights of the neurons are initialized to small random values. The network is presented with a large number of data vectors and the weights are adjusted using a competitive learning algorithm. When each data vector is presented to the network, the competitive learning algorithm calculate the Euclidean distance to all weight vectors. The neuron with the weight vector most similar to the input data vector is identified as the "winner". The weights of the "winner" neuron and the weights in its neighborhood are adjusted such that

they become more similar to the input vector. At the end of the training procedure the input neurons are evenly distributed over the training data with each output neuron representing a particular input data pattern most closely. During the mapping phase, as each input pattern is presented to the network, a single winning neuron, whose weight vector lies closest to the output vector, will be determined. Typically, these winning neurons are labeled or colored and the resulting pattern is what is referred to as the "map". This map can be used much in the same way as PCA to determine if there is some latent structure to the groupings represented by the map. Or the SOM can be used as a tool for further classification. To use the SOM as a classification or pattern recognition tool (this is the testing phase), novel input data patterns are presented to the network and in principle, the input dat patterns can be classified according to what regions of the SOM are "activated".

Figure 3.5: Self Organizing Map



Chapter 4

Quantum Mechanically Derived Descriptors in Mono-Substituted Benzenes

4.1 Introduction

Benzene, C_6H_6, is the archetypal aromatic hydrocarbon, with a continuous cyclic array of pi bonds with a D6h planar geometry. This particular structure results in carbon-carbon bond lengths that are intermediate between double and single bond, which is consistent with electron delocalization. In this way, the structure exists as a superposition of resonance structures, rather than one single form. This special character of benzene imparts thermodynamic stability compared with other aromatic compounds, contributing to the peculiar molecular and chemical properties associated with its reactivity. When one of the hydrogens of the benzene ring is substituted with another functional group, this changes the reactivity, either making it slower or faster. If the substituent makes a reaction slower, then it is said to deactivate the ring; if the substituent makes a reaction faster, it is said to activate the ring. Substituents also result in very specific products, classified as ortho-, meta-, or para-, substituted products (Figure 4.1), leading to a very specific classification as shown in the table.

In Category I, the substituent directly attached to benzene is more electronegative than the carbon it is attached to, meaning it has an inductive electron withdrawing effect (-I; except for O-). Countering this effect, is a strong resonance electron donating effect (+Re) due to the lone pairs the group can donate to the ring through resonance contribution (Figure 4.2). This category of substituent makes the benzene ring more electron rich, and thereby more susceptible to attack by an electrophile (activating groups). The places where charge is built up in the resonance forms shows where the directing group orientation will be. In these cases, one sees that a negative charge is built up only on the ortho and para positions, and therefore one would expect that an electrophile would attack those positions selectively.

In Category II, because there is a full/partial positive charge on the element of the functional

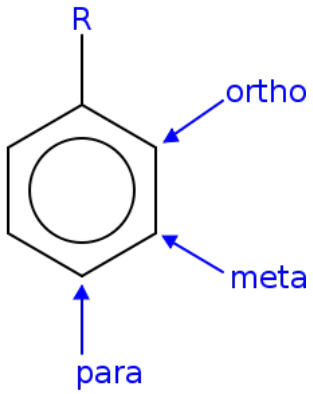


Figure 4.1: Ortho, Meta and Para Positions on benzene

Table 4.1: Substituted Benzenes

effect	Category	Type	Substituent
activating ortho/para directing	Category I	Inductive EWG (-I) Resonance EDG (+Re)	O-, NR_2, NHR, NH_2 OH, OR, NHCOR, OCOR
strongly deactivating meta directing	Category II	Inductive EWG (-I) Resonance EWG (-Re)	NO_2, CN, SO_3H, CHO COR, CO_2H, $CONH_2$
mild activating ortho/para directing	Category III	weak Resonance EDG (+Re)	alkyl (R)
deactivating meta directing	Category IV	strong Inductive EWG (-I)	NH_3, CF_3
deactivating ortho/para directing	Category V	strong Inductive EWG (-I) very small (+Re)	halogens (Cl, F, Br, I)

group attached to the carbon of benzene, they all have moderate to strong EWG inductive effects (-I). Additionally, they all have resonance EWG character (-Re) as shown in Figure 4.3. In this way, they make the benzene ring more electron poor (deactivate), causing reactions to proceed much slower. The resonance structure show that these functional groups cause the benzene to be meta directing.

In Category III, there is no electronegativity difference effect, since the substituent directly attached to the benzene ring is also a carbon species. However, there are weak resonance effects that increase the electron density in the ring through hyperconjugation (+Re), which is a relatively weak effect making these functional groups weakly activating, and also ortho-, para-, directing.

Category IV includes functional groups that have strong electron-withdrawing inductive tendencies (-I), either due to their positive charge, or due to highly electronegative halogen atoms (Figure 4.4)There is no resonance effects through orbital interaction nor through electron pair

Figure 4.2: Category I

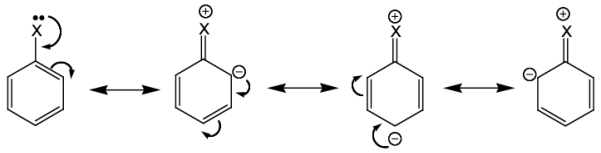


Figure 4.3: Category II

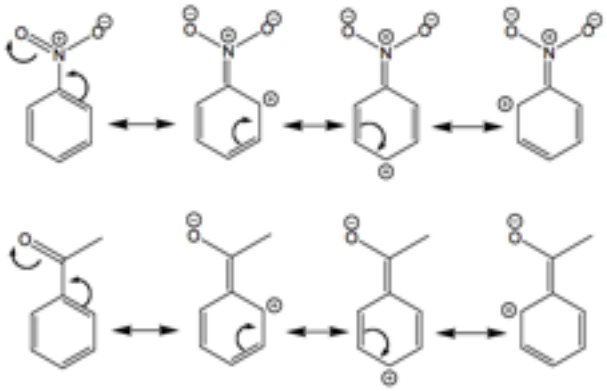
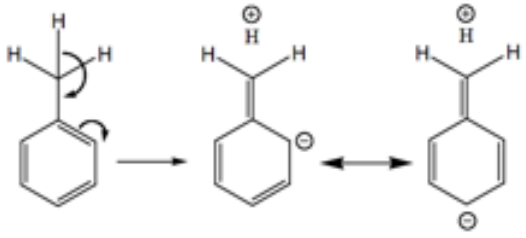


Figure 4.4: Category III



resonance to the benzene pi system. As such, these groups are deactivating and meta directing.

Category V is comprised of the halogens. The electronegativity of the halogens will withdraw electron density from the ring via a strong inductive effect and the lone pair will donate electron density via a weak resonance effect. The inductive effect deactivates the ring but the resonance will cause the benzene to be ortho-para directing

With more than one substituent on a benzene ring, one can determine directing abilities if the groups direct in a similar manner thereby reinforcing a particular directing tendencies. However, if the substituents have opposing directing tendencies then it can become more complicated to determine the outcome of a reaction. There are, however, certain conventions used for determination of directing tendencies of multiple substituents, however. Typically, ortho-, para- directors have more influence than meta- directors, and of the former, there is the following hierarchy $O-, NR_2, NHR, NH_2, OH, OR > NHCOR, OCOR > R > F, Cl, Br, I >$ meta directing groups. Steric considerations also matter quite a lot, for example, typically reaction will not occur at the site when there are already substituents on either side of that position.

The electron density distribution relates the local and global properties of a molecule in accord with fundamental theorems of molecular physics. These relations can be used to extract physical meaning from similarities found between calculated molecular properties and an individual functional groups, in addition to the spatial requirements and global shape of the molecule.[7] Exploring the connection between chosen descriptors (e.g., Chapter 2) and their relation to the physicality of the molecule, is the goal of this chapter. This will be demonstrated by evaluating whether or not key descriptors explain what is known about mono-substituted benzenes from basic chemical theory.

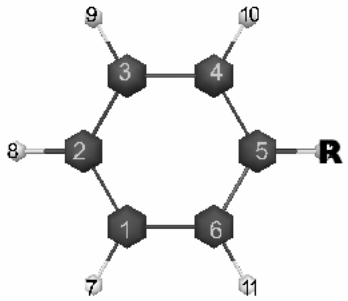
Table 4.2: Substituents Used in this Study

Substituents	
Effect	Substituent
Category I	OH, NH_2, OCH_3
Category II	CN , NO_2
Category III	CH_3
Category IV	CF_3 , NH_3
Category V	Br

4.2 Data-set and Computational Details

The numbering scheme used for the set of substituted benzenes in this study is shown in Figure 4.5 where R represents the substituent. The list of substituents is shown in Table 4.2. All calculations have been carried out using the GAMESS software package.[1] All geometries were optimized using Restricted Hartree-Fock (RHF) method with the DZV(2d,p) basis set. For the screening charge density calculations, the BP86/KTZP(2d,p) level of theory was used. The descriptor calculations were performed using the custom software described in Chapter 8. In this chapter, only results for substituent OH, as an example of a strong ortho-para director, and NO_2, as an example of a strong meta director, are discussed to exemplify the utility of these particular descriptors.

Figure 4.5: Numbering Scheme for Mono-Substituted Benzenes



4.3 Results and Discussion

4.3.1 Electron Density based Descriptors

The electron density was calculated at a single point, 0.7 Bohr directly above the nuclei of the 6 Carbon atom horizontal plane of the D6h benzene structure. To get a complete picture of the π electron withdrawing or donating effect of the substituent, the electron density should be sampled at a regular grid in the plane rather than a single point, however for the purposes of this study the goal was to evaluate if a single point would suffice as a descriptor.

For the Category I substituents (-OH and -OCH_3) the electronegativity of the substituent will have a small inductive electron withdrawing effect on the ring, and the lone pairs on the oxygen will contribute a strong electron donating effect through resonance. The resonance results in electron density being directed to ortho and para positions of the ring. This results in these sites being more nucleophilic, or in other words the resulting functionalized benzene is reactive towards electrophiles at the ortho and para sites. A density differences between the functionalized bezene and an unsubstituted parent benzene indeed shows an increase in electron density at the ortho and para positions, as well as a more moderate decrease in electron density at the meta positions, as shown in Table 4.3. The Category II substituents (NO_2 and CN) should have a strong inductive electron withdrawing effect as well as an electron withdrawing resonance effect, which results in a depletion of electron density at the ortho and para positions. As expected, there is a slight decrease in electron density at the ortho and para positions with a slight increase in electron density at the meta positions. The sigma electrons of the C-H bonds in the methyl of the Category III substituent (CH_3) with the adjacent π orbitals of the ring allow an increase in electron density of the ring via a weak resonance effect and directs electron density to the ortho-para positions. The meta directing pattern is not as clearly seen in the Category IV substituents CF_3 and NH_3. These substituents have strong electron withdrawing inductive tendencies stemming from the highly electronegative Fluorine atom in the case of the CF_3 substituent and from the positive charge of the ammonia group but the meta pattern is not evident from the electron densities taken at a single point 0.7 Bohr above the plane of the ring. A more comprehensive sampling of the π electron density would be required in this scenario to gauge the overall electronic effects of Category IV substituents. The electronegative Category V substituent (Br) will withdraw electron density from the ring via a strong inductive effect and the lone pair will donate electron density via a weak resonance effect which will direct the electron density ortho-para.

4.3.2 Molecular Orbital based Descriptors

An alternative to evaluating the total electron density in describing reactivity, is to think about the localization of the HOMO and LUMO orbitals. As the highest energy occupied and lowest energy unoccupied orbitals, the electrons in these orbitals are the most available to participate in a chemical reaction. As was described in Chapter 2, descriptors based on molecular orbitals can be used as indicators of reactivity. As the energy of the HOMO is related to ionization potential, it can characterize how susceptible a molecule would be to attack by an electrophile. A harder nucleophile would have a lower energy HOMO, since the electrons in this orbital would be less energetically accessible. Similarly, the LUMO energy is related to electron affinity and is indicative of the susceptibility to attack by nucleophiles with softer electrophiles having lower LUMO energies since these orbitals would be less energetically accessible to incoming electrons. The orbital density (square of the eigenvector coefficients) is also an important indicator of reactivity since a majority of chemical reactions take place at the position and in the orientation where the overlap of the HOMO and LUMO can reach a maximum. For electron donors, the HOMO density is important while for electron acceptors it is the LUMO density that is important.

One approximation made in the determination of the molecular orbital (MO) descriptor (see, e.g., Chapter 2) involves the sum of the atomic orbital coefficients of the atom they are centered on. This is not strictly rigorous however, since the contribution of each atomic orbital involved in the HOMO or LUMO orbitals are not strictly localized to a particular atom, but often span a large portion of a molecule, making the interpretation more difficult. Reactivity is not always completely determined by the frontier orbitals. For example, if the HOMO envelopes an aromatic system, it is less likely to react as a nucleophile because this will lead to loss of resonance stabilization. Alternatively, in some molecules, reactivity is determined by strong electrostatic interactions, rather than by frontier orbital overlap.

It is well-known that benzene itself is sufficiently nucleophillic to undergo electrophilic aromatic substitution by acylim ions or alkyl carbocations to give substituted derivatives. Typically the effect of substituents on a benzene ring are discussed in terms of their effect on the rate and orientation of further electrophilic aromatic substitution reactions. In the following discussion, both nucleophilicity and electrophilicity are discussed synchronously as a basis for comparing how well the known electron donating/withdrawing effects of the substituents are reflected in the descriptors.

Global MO Descriptors: HOMO, LUMO, HOMO-LUMO gap, Sums of atom-based MO descriptors

The energy of the HOMO and LUMO as well as the magnitude of the HOMO-LUMO gap are frequently used to describe or predict reactivity between molecules. The molecular orbital based descriptors (frontier orbital density and superdelocalizabilities) described in Chapter 2

Table 4.3: Density Difference from Benzene (density_{subs} - $\text{density}_{benzene}$) / $\text{density}_{benzene}$

atom	benzene	electron density difference								
atom	H	OH	NO_2	NH_3	NH_2	Me	MeO	CN	CF_3	Br
c1:meta	0.2147	-1.52%	1.23%	-1.39%	-2.04%	-0.51%	-1.32%	0.73%	0.62%	-0.14%
c2:para	0.2147	3.74%	-2.58%	-2.65%	4.12%	1.05%	3.55%	-1.69%	-1.23%	0.33%
c3:meta	0.2147	-1.93%	1.24%	-1.20%	-2.02%	-0.54%	-1.57%	0.74%	0.50%	-0.14%
c4:ortho	0.2147	5.20%	-1.89%	1.84%	6.32%	1.78%	4.90%	-0.74%	0.01%	1.96%
c5:R	0.2147	-0.32%	12.80%	15.72%	-3.06%	-0.95%	-0.06%	7.56%	6.68%	8.60%
c6:ortho	0.2147	6.94%	-1.88%	2.10%	6.34%	1.60%	6.03%	-0.74%	-0.61%	1.96%

can also be used to describe reactivity of individual atoms or they can be summed to describe the reactivity for fragments or entire molecules. Tables 4.4 and 4.5 list all the global descriptors based on the occupied and unoccupied molecular orbitals respectively while Table 4.6 lists the descriptors based on the HOMO-LUMO gap as well as the dipole. The HOMO-LUMO gap ranking is in Table 4.6.

In Table 4.4, the strongly activating Category I substituents have the highest HOMO values as expected since these substituents donate electron density to the ring. This electron density makes the ring generally softer. Soft nucleophiles have a higher energy HOMO since these electrons are susceptible to attack by an electrophile. The strongly deactivating electron withdrawing groups of Category II and IV have the lowest energy HOMO which corresponds to their behavior as relatively harder nucleophiles since the lower lying HOMO is less susceptible to attack. The more mildly activating/deactivating groups are placed in the middle and follow the expected ranking: NH2 > MeO > OH > Me > Br > H > CF_3 > CN> NO_2 > NH_3. The sum of the electrophilic frontier orbital density describes the density of the orbitals rather than the energetics so it is not surprising that there is not a trend that reflects the relative energetics. The sum of the approximate superdelocalizability however, does include the respective energy eigenvalue of the orbital (the density of the orbital divided by the energy eigenvalue) and the expected trend is again seen. The sum of the electrophilic superdelocalizability shows expected trend from most strongly activating to most strongly deactivating as was the case for the HOMO with the exception of toluene.

The descriptors based on the unoccupied orbitals also are ranked in Table 4.5. As was the case for the HOMO, the LUMO also shows the expected trend with the strongly deactivating Category II substituents having the highest LUMO energy, which means that their orbitals are "softer" and more accessible to incoming electrons (attack by a nucleophile). The sum of the nucelophilic frontier orbital density better represents the general trend than was the case for the electrophilic frontier orbital density, with the strongly activating/deactivating groups at the opposite ends of the table. AS was the case for the corresponding electrophilic descriptor, the approximate nucleophilic superdelocalizability also better represents the overall trend with the exception of Me and NH_3. The charged ammonia group is an exception to the trend due the influence of its positive charge on the orbital energetics. Unlike the descriptors based on the unoccupied orbitals, the nucleophilic superdelocalizability does not reflect the expected trends based on electron withdrawing or donating ability.

Table 4.6 lists these variables in the order of increasing HOMO-LUMO gap. The HOMO-LUMO gap is also used as a stability index and is directly proportional to the hardness. The relatively unreactive unsubstituted parent system has the largest HOMO-LUMO gap. It is expected that the relative stability of the unsubstituted benzene would be in between the activating and deactivating substituents (as was the case for the HOMO and electrophilic superdelocalizability). The exact ordering of the substituents generally reflects the expected stability trends

Table 4.4: Global(summed over 6 C atoms in benzene) occupied MO based descriptors (ranked from least to greatest)

	HOMO		elecFOD		approxElecSuper		elecSuper
NH_3	-0.502	CF_3	0.872	H	-6.484	NH_2	-84.596
NO_2	-0.3669	Br	0.889	OH	-3.362	MeO	-80.863
CN	-0.359	CN	0.972	NH_2	-3.334	OH	-79.664
CF_3	-0.3543	Me	0.974	MeO	-3.309	Br	-75.876
H	-0.3341	NH_2	0.979	Me	-3.027	H	-74.046
Br	-0.3316	MeO	1.018	NO2	-2.938	CN	-69.898
Me	-0.3218	OH	1.050	CN	-2.706	NO_2	-68.448
OH	-0.3123	NO_2	1.078	Br	-2.681	Me	-64.138
MeO	-0.3077	H	1.083	CF_3	-2.461	CF_3	-58.929
NH_2	-0.2937	NH_3	1.089	NH_3	-2.169	NH_3	-49.740

but of course the HOMO-LUMO gap is affected by other electronic and steric features of the molecule so the results do not strictly follow the trends found in experimental data for the relative rates. After benzene, the next two "hardest" substituents are the deactivating Category IV substituents (CF_3 and NH_3+) which have a strongly inductive electron withdrawing effect with no resonance. This decreases the electron density of the π system in the ring thus increasing the hardness. On the other hand, the strongly deactivating Category II nitro and cyano groups lie at the other extreme end of the table. These also have have a strong inductive electron withdrawing effect as well as an electron withdrawing resonance character which destabilizes the π electrons in the benzene ring making the molecule "softer". The strongly activating Category I substituents, MeO, NH_2 and OH, donates π electrons via a strong resonance effect making the molecules softer. This is countered by a weaker inductive electron withdrawing effect. The relative stabilities of Bromobenzne and toluene provides a good example of how the relative ordering of these substituents is effected by a balance of resonance and inductive effects. The weakly activating Category III methyl substituent does not have any inductive capability but donates π electrons to the sigma frame though weak resonance effect, while the Category V bromo substituent strongly withdraws electron density via induction and only donates electron density to the π framework via a weaker resonance effect. The net effect is that bromobenzene is "softer" than toluene due to the additional density in the π cloud even though the experimental data for relative rates of nitration indicate that the nitration of toluene proceeds at a much faster rate.

Table 4.5: Global(summed over 6 C atoms in benzene) unoccupied MO based descriptors (ranked from least to greatest)

	LUMO		nucFOD		approxNucFOD		nucSuper
NH_3	-0.0425	NO_2	0.737	H	-24.107	H	-1434.893
NO_2	0.0542	CN	1.151	CN	-14.755	CF_3	-1431.151
CN	0.078	Me	1.295	NO_2	-13.603	Me	-1381.441
CF_3	0.1036	NH_3	1.322	CF_3	-13.227	MeO	-1377.150
Br	0.1146	CF_3	1.370	Br	-13.197	OH	-1316.642
OH	0.1262	Br	1.512	OH	-12.390	NH_3	-1214.646
MeO	0.1301	MeO	1.561	NH_2	-12.217	CN	-1194.654
Me	0.1303	OH	1.564	MeO	-12.001	NH_2	-1165.933
H	0.1308	H	1.577	Me	-9.939	NO_2	-1072.160
NH_2	0.1321	NH_2	1.614	NH_3	31.107	Br	-177.909

Table 4.6: Global MO descriptors (in Hartrees) and Dipole Moment (Debye) ranked according to decreasing HOMO/LUMO gap

substituent	HOMO	LUMO	HOMO/LUMO gap	hardness	softness	dipole
H	-0.3341	0.1308	0.4649	0.23245	2.1510	0.0000
NH_3	-0.502	-0.0425	0.4595	0.22975	2.1763	7.3504
CF_3	-0.3543	0.1036	0.4579	0.22895	2.1839	2.8761
Me	-0.3218	0.1303	0.4521	0.22605	2.2119	0.3394
Br	-0.3316	0.1146	0.4462	0.2231	2.2411	1.9736
OH	-0.3123	0.1262	0.4385	0.21925	2.2805	1.3518
MeO	-0.3077	0.1301	0.4378	0.2189	2.2841	1.3040
CN	-0.359	0.078	0.437	0.2185	2.2883	4.8884
NH_2	-0.2937	0.1321	0.4258	0.2129	2.3485	1.4752
NO_2	-0.3669	0.0542	0.4211	0.21055	2.3747	4.9181

Localized Descriptors

Frontier orbital densities have been used on atoms to characterize donor acceptor interactions.[8] One limitation of the frontier orbital densities is that they can only be used to describe the reactivity of different atoms on the same molecule. To compare reactivities across different molecules the frontier orbital density needs to be normalized by dividing the density by the corresponding energy of that orbital. This normalized frontier orbital density descriptor is also known as the approximate superdelocalizability. The normalized (shown in Tables 4.7 and table:elecFOD) and un-normalized (Tables 4.9 and 4.10) frontier orbital densities for this monosubstituted benzene dataset clearly demonstrate ortho-meta-para directing trends for the strongest Category I and II substituents (e.g. OH in Category I and NO_2 in Category II). For phenol, both the normalized and un-normalized electrophilic frontier orbital densities are highest at the ortho and para positions, while the nucleophilic frontier orbital density is lowest at the ortho and para positions and the opposite is seen for nitrobenzene. This corresponds to what is known about the reactivity of phenol; the ortho and para positions on the ring have increased electron density and are therefore the sites of electrophilic attack by an incoming nucleophile. Similarly for nitrobenzene, the meta positions have the largest magnitudes for the electrophilic descriptors at the meta positions. The ortho-meta-para patterns are less clear for the substituents that more weakly activate or deactivate the ring. Also, the meta directing substituents which have less influence on the ring than the ortho-para directors also do not show the pattern as clearly as the ortho-para directors. The nucleophilic frontier orbital densities (based on the unoccupied LUMO orbitals) do not show the patterns as clearly as the electrophilic densities (based on the occupied HOMO orbitals) due to the occupied orbitals being better represented computationally.

The data can be better understood by looking at a visual of the HOMO and LUMO molecular orbitals (Figures 4.6 - 4.7), particularly since the coefficients of these orbitals are used in the expressions for the electrophilic and nucleophilic frontier orbital densities, respectively. The MO surfaces shown were all taken a contour value of 0.100. The blue/red color of the orbitals of the phase of the lobes is irrelevant here since the equation for frontier orbital density represents the phase of the particular orbital, calculated as the square of the coefficients. In the case of nitrobenzene, the HOMO shows density primarily at the meta and ortho positions. Since the total density is partitioned into all occupied molecular orbitals, the HOMO and LUMO show only the density relevant to that particular energy level.

Sometimes it is advantageous to look at more than just the frontier orbitals to fully determine reactivity. This is what is done in the "superdelocalizability" descriptor. Superdelocalizability incorporates all of the occupied or unoccupied molecular orbitals in a molecule, as opposed to just the HOMO and the LUMO in the computation of the electrophilic and nucleophilc superdelocalizabilities respectively. This could lead to a better representation of the reactive space, however, If the electrons in the frontier orbital dominate the interactions between the

Figure 4.6: HOMO for phenol nitrotoluene

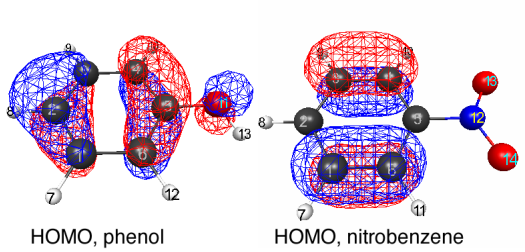
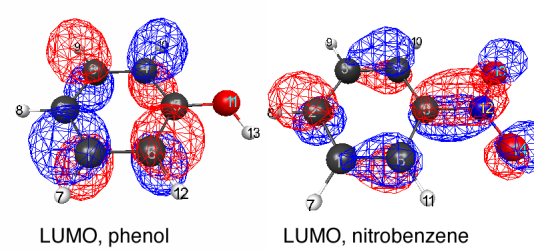


Figure 4.7: LUMO for phenol and nitrobenzene



acceptor and donor taking part in the reaction, then the superdelocalizability descriptor could add more noise to the data making interpretation more difficult. This is what is seen here with using The electrophilic superdelocalizability(Tables 4.11) represents the expected directing patterns for the ortho and para directors fairly well but does not represent the meta directors as clearly. As was the case for the previous descriptors based on the unoccupied orbitals, the nucleophilic superdelocalizability (Table 4.12) do not represent the patterns as well as the electrophilic analog.

Table 4.7: Electrophilic Frontier Orbital Densities (MP2/DZV(2d,p))

atom	H	OH	NO_2	NH_3	NH_2	Me	MeO	CN	CF_3	Br
c1:meta	0.181	0.084	0.272	0.217	0.049	0.102	0.100	0.083	0.313	0.061
c2:para	0.181	0.361	0.002	0.003	0.335	0.351	0.348	0.306	0.013	0.285
c3:meta	0.181	0.045	0.271	0.245	0.049	0.057	0.030	0.083	0.216	0.061
c4:ortho	0.181	0.160	0.265	0.297	0.155	0.131	0.169	0.077	0.318	0.099
c5:R	0.181	0.267	0.003	0.005	0.237	0.334	0.250	0.346	0.011	0.284
c6:ortho	0.181	0.133	0.265	0.322	0.155	0.094	0.121	0.077	0.212	0.099
sum	1.083	1.050	1.078	1.089	0.979	0.974	1.018	0.972	0.872	0.889

Table 4.8: Nucleophilic Frontier Orbital Densities (MP2/DZV(2d,p))

atom	H	OH	NO_2	NH_3	NH_2	Me	MeO	CN	CF_3	Br
c1:meta	0.263	0.420	0.029	0.012	0.402	0.503	0.454	0.070	0.222	0.385
c2:para	0.263	0.002	0.237	0.272	0.004	0.077	0.012	0.388	0.454	0.001
c3:meta	0.263	0.386	0.029	0.253	0.404	0.198	0.344	0.070	0.026	0.385
c4:ortho	0.263	0.373	0.124	0.012	0.384	0.474	0.420	0.142	0.256	0.369
c5:R	0.263	0.002	0.195	0.343	0.035	0.044	0.006	0.339	0.413	0.002
c6:ortho	0.263	0.381	0.124	0.430	0.386	0.201	0.327	0.142	0.048	0.369
sum	1.577	1.564	0.737	1.322	1.614	1.295	1.561	1.151	1.370	1.512

Table 4.9: Approximate Superdelocalizabilities (Normalized Electrophilic Frontier Orbital Densities) (MP2/DZV(2d,p))

atom	H	OH	NO_2	NH_3	NH_2	Me	MeO	CN	CF_3	Br
c1:meta	-1.081	-0.270	-0.740	-0.432	-0.165	-0.317	-0.324	-0.231	-0.884	-0.183
c2:para	-1.081	-1.157	-0.006	-0.006	-1.142	-1.089	-1.130	-0.852	-0.038	-0.860
c3:meta	-1.081	-0.145	-0.740	-0.489	-0.166	-0.177	-0.097	-0.232	-0.610	-0.183
c4:ortho	-1.081	-0.512	-0.722	-0.591	-0.526	-0.406	-0.550	-0.214	-0.897	-0.299
c5:R	-1.081	-0.854	-0.008	-0.009	-0.808	-1.038	-0.812	-0.963	-0.032	-0.858
c6:ortho	-1.081	-0.424	-0.722	-0.642	-0.527	-0.292	-0.395	-0.214	-0.599	-0.299
sum	-6.484	-3.362	-2.938	-2.169	-3.334	-3.027	-3.309	-2.706	-2.461	-2.681

Table 4.10: Normalized Nucleophilic Frontier Orbital Density (MP2/DZV(2d,p))

atom	H	OH	NO_2	NH_3	NH_2	Me	MeO	CN	CF_3	Br
c1:meta	-4.018	-3.331	-0.529	0.290	-3.040	-3.859	-3.486	-0.895	-2.144	-3.362
c2:para	-4.018	-0.017	-4.366	6.408	-0.027	-0.588	-0.090	-4.976	-4.379	-0.013
c3:meta	-4.018	-3.057	-0.529	5.947	-3.057	-1.519	-2.640	-0.896	-0.248	-3.362
c4:ortho	-4.018	-2.953	-2.290	0.280	-2.908	-3.639	-3.227	-1.819	-2.468	-3.222
c5:R	-4.018	-0.014	-3.599	8.063	-0.265	-0.334	-0.046	-4.352	-3.989	-0.016
c6:ortho	-4.018	-3.019	-2.290	10.120	-2.920	-1.540	-2.511	-1.818	-0.464	-3.222
sum	-24.107	-12.390	-13.603	31.107	-12.217	-9.939	-12.001	-14.755	-13.227	-13.197

Table 4.11: Electrophilic Superdelocalizabilites (MP2/DZV(2d,p))

atom	H	OH	NO_2	NH_3	NH_2	Me	MeO	CN	CF_3	Br
c1:meta	-12.341	-13.181	-11.325	-8.211	-13.985	-12.796	-13.395	-11.574	-11.696	-12.481
c2:para	-12.341	-13.319	-11.272	-8.253	-14.182	-12.863	-13.492	-11.559	-11.683	-12.490
c3:meta	-12.341	-13.179	-11.325	-8.211	-13.985	-12.807	-13.386	-11.574	-11.701	-12.481
c4:ortho	-12.341	-13.465	-11.517	-8.390	-14.357	-13.009	-13.701	-11.756	-11.950	-12.774
c5:R	-12.341	-12.901	-11.493	-8.283	-13.730	-12.663	-13.128	-11.680	-11.900	-12.876
c6:ortho	-12.341	-13.619	-11.517	-8.391	-14.357	-13.017	-13.761	-11.756	-11.971	-12.774
sum	-74.046	-79.664	-68.448	-49.740	-84.596	-64.138	-80.863	-69.898	-58.929	-75.876

Table 4.12: Nucleophilic Superdelocalizabilites (MP2/DZV(2d,p))

atom	H	OH	NO_2	NH_3	NH_2	Me	MeO	CN	CF_3	Br
c1:meta	-238.757	-199.686	-148.539	-174.367	-170.707	-247.875	-216.042	-167.436	-239.158	-25.050
c2:para	-240.006	-201.271	-148.084	-178.254	-170.125	-246.071	-200.985	-167.484	-235.053	-25.239
c3:meta	-238.678	-198.266	-148.463	-175.947	-170.558	-248.614	-199.346	-167.365	-239.227	-25.050
c4:ortho	-238.753	-231.178	-183.939	-213.226	-202.913	-296.176	-238.257	-205.218	-305.305	-30.543
c5:R	-240.011	-257.296	-259.283	-263.574	-249.065	-342.705	-270.401	-282.034	-412.409	-41.485
c6:ortho	-238.687	-228.945	-183.852	-209.279	-202.565	-296.069	-252.120	-205.117	-302.182	-30.543
sum	-1434.893	-1316.642	-1072.160	-1214.646	-1165.933	-1381.441	-1377.150	-1194.654	-1431.151	-177.909

Table 4.13: CHELPG benzene gp and solvated

	gp	solvated
c1:meta	-0.1901	-0.2139
c2:para	0.0625	0.0487
c3:meta	-0.1901	-0.2138
c4:ortho	-0.1902	-0.2139
c5:R	0.0626	0.0487
c6:ortho	-0.1902	-0.2139

4.3.3 Partial Charge

The partial charge describes the build-up or depletion of charge on individual atoms. In benzene, every carbon atom has the same electronegativity so it is expected that the partial charge distribution would be identical among Carbon atoms. However this is not what is observed when the CHELPG charges are calculated as seen in Figure 4.8. Problems in assigning charge values have been observed[9] with polar donor-acceptor charge complexes when using electrostatic potential-based charge methods (including CHELPG) since these methods assign charge based on what the molecule "feels" as the probe charge approaches and the fitting scheme does not always partition the charge perfectly. This phenomenon is typically more of a problem for charged push-pull systems and was not expected to be a problem for benzene. Nevertheless, the ortho-meta-para directing pattern for benzeneOH can be seen from the CHELPG charges with the ortho and para positions having a negative partial charge and the meta positions having a positive partial charge. Likewise for nitrobenzene the meta positions are the most negative. The results are less clear for nitro though as the ortho positions are also negative, albeit less so.

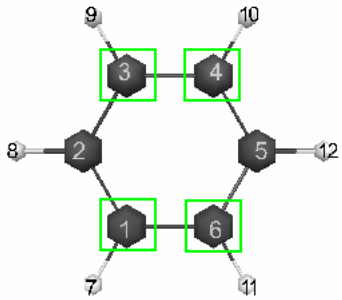


Figure 4.8: CHELPG Pattern for unsubstituted benzene. Values are are equal for 1,3,4,6 and 2,5

Table 4.14: CHELPG phenol and nitrotoluene

atom	OH	NO_2
c1:meta	0.0640	-0.2244
c2:para	-0.3225	0.1137
c3:meta	0.0572	-0.2224
c4:ortho	-0.3701	-0.1656
c5:R	0.5192	0.1778
c6:ortho	-0.4409	-0.1794

4.3.4 Solvent screening Charge Density as a Fingerprint

To investigate the effect of solvent environment on the set of mono-substituted benzenes, calculations where carried out on the molecules in a water environment and the charges representing the interaction of the electronic density of the molecule with a continuum dielectric of 80, was evaluated. The full details of the implementation of the screening charge density profile algorithm and accompanying data structures are given in Chapter 8, the Computational Methods chapter. Here, the focus will be on the scientific results. Since the screening charge density is an indicator of overall polarity of a molecule, one does not expect that same indicators of reactivity as the above descriptors, in terms of ortho-meta-para directing patterns. However, the underlying inductive and resonance effects responsible for the patterns associated with the substituents are still important.

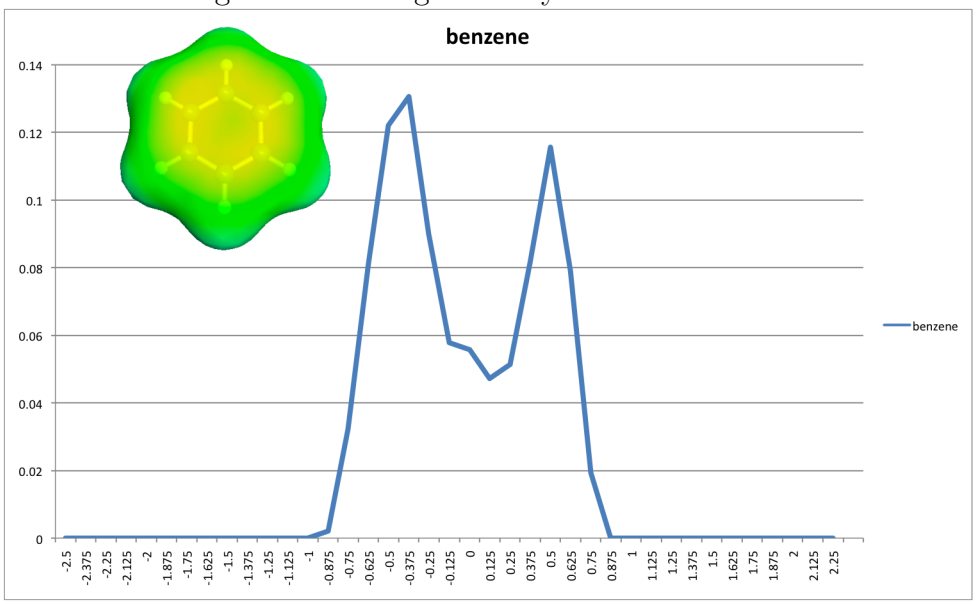
In the graphs of solvent charge response (fingerprint), there is a sign inversion of the polarization charge density σ compared to molecular polarity, since these values represent the response of the molecule to the water environment. For example, if there is significant negative charge localization in the molecule, one expects a response with a complimentary positive charge in the solvent environment. The area on the left between -2.5 e/A^2 until -1 e/A^2, is sometimes referred to as the H-bond donor region, since this area is associated with positively charged species and electrophillic tendencies. The area on the right between 1 e/A^2 and 2.5 e/A^2, is sometimes referred to as the H-bond acceptor region, since this area is associated with negatively charged species and nucleophillic tendencies. The center region, is the non-polar region.

In the specific case of benzene, there are two main peaks symmetrically places at -0.6 and +0.6 e/A^2. the peak at -0.6 in benzene corresponds to the very small negative charge density of the delta positive belt of hydrogens around the ring, and the peak at +0.6 corresponds to the small delta minus charge associated with the π face (quadrupolar region) of the ring.

Category I substituents

The screening charge density histograms for -OH, -NH_2 and -OMe substituted benzene, are representative of Category I substituents. These three substituents all possess lone pairs that will

Figure 4.9: Charge Density Profile Benzene

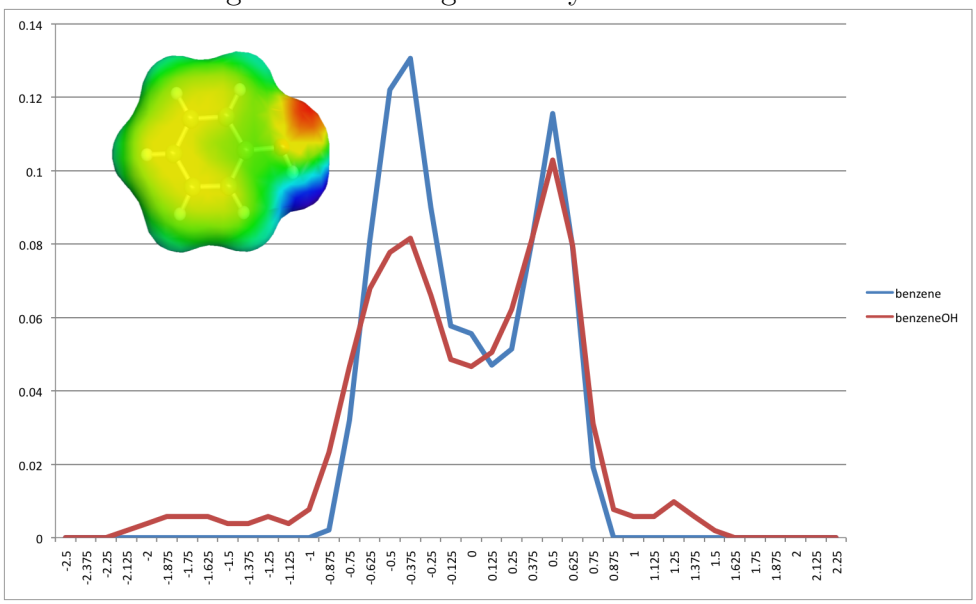


donate electron density to the ring through resonance. The electronegativity of the substituent will also draw some amount of electron density from the ring as well as provide a negative partial charge. The partial negative charge of the substituent is represented in all three cases by the small lump at +1 e/A^2. The peak on the left is associated with positive charge density (and electrophilic behavior) and in all three cases it decreases in height. This means that there is less surface area of the molecule associated with positive charge. This could be interpreted as meaning these substituents make the positively charged hydrogens less electrophilic, which would be expected since these groups have a strong electron donating effect. It was also expected that the peak associated with the negatively charged π face would change since the electron density of the π face is responsible for the orho-meta-para patterns evidenced in the previous section. However, this peak did not change significantly (compared to the peak associated with the positively charged hydrogens). This could be because the ortho-meta-para directing patterns represent a change in the distribution patterns but the overall charge density of the π

The screening charge density histograms for -OH, -NH_2 and OMe substituted benzene, are representative of Category I substituents. Compared to the parent system, all three substituents have peaks in approximately the same place as for unsubstituted benzene, but with varying peak heights. These three substituents all possess lone pairs that donate into to the ring through resonance. The strong electronegativity of the substituent in turn results in a strong inductive electron withdrawing effect through the sigma system.

The solvent surface response for phenol is depicted in Figure 4.10. The substituted systems show significant perturbation of the benzene peak representing the partial positive belt around the ring; in phenol, this peak is now decreased as the substituent changed the nature of that belt area. Instead, there are two small peaks at approximately -1.5 and + 1.2 nm . The peak

Figure 4.10: Charge Density Profile Phenol



at -1.5 is in the 'H-bond donor region' and corresonds to the hydrogen of the subsituent, and the peak at +1.2 is in the h-bond acceptor region and is associated with the strongly negative polar regions of the oxygen lone pairs. As such, phenol would have improved ability to form hydrogen bonds over unsubstituted benzene. In general, one sees additional small regions on the left side of the spectrum, associated with more strongly positive charge regions around the ring compared to benzene. The peak associated with the negatively charged π face did not change significantly, indicating that the partial negative pi face is of similar magnitude and is across a similar amount of the pi surface. Unlike the ortho-meta-para directing patterns associated with the electron density, the present patterns represent a response of that electron density to the solvent environment in terms of charge density, so one can not expect to actually see focusing of charge to these regions.

NH_2 is also a strongly activating electron donating group with a lone pair on the N of the substituent group. The peak in the electrophilic region (at -0.6 nm) is again slightly smaller than benzene, but larger than phenol (Figure 4.12). In this case, multiple peaks appear further in the negative region, associated with the hydrogens of the substituent. Interestingly, the effect of the NH2 on the rings pi cloud is more significant than observed in phenol, showing an increase in partial negative charge in the pi region compared to benzene, but over a smaller region of the π surface.

The anisole charge density profile looks very similar to benzene. The peak associated with the positively charged hydrogens being only slightly greater in anisole than in benzene. Since anisole has less electron donating ability than phenol and aniline, the relative trend corresponds to what is expected. The peak in the nucleophilic region that corresponds to the negative pi face is also slightly greater in anisole. This result is also expected since the methoxy group

Figure 4.11: Charge Density Profile BenzeneNH_2

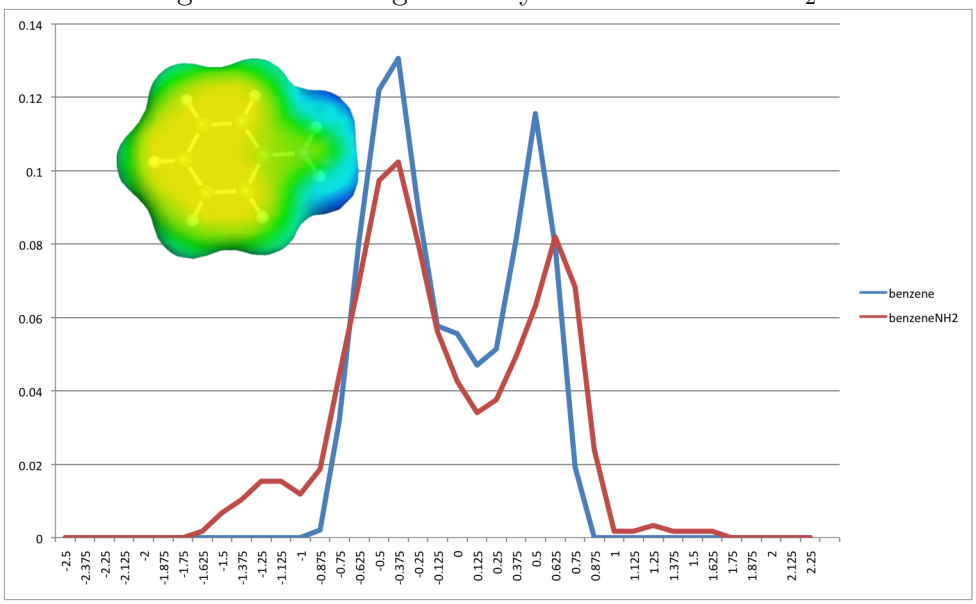
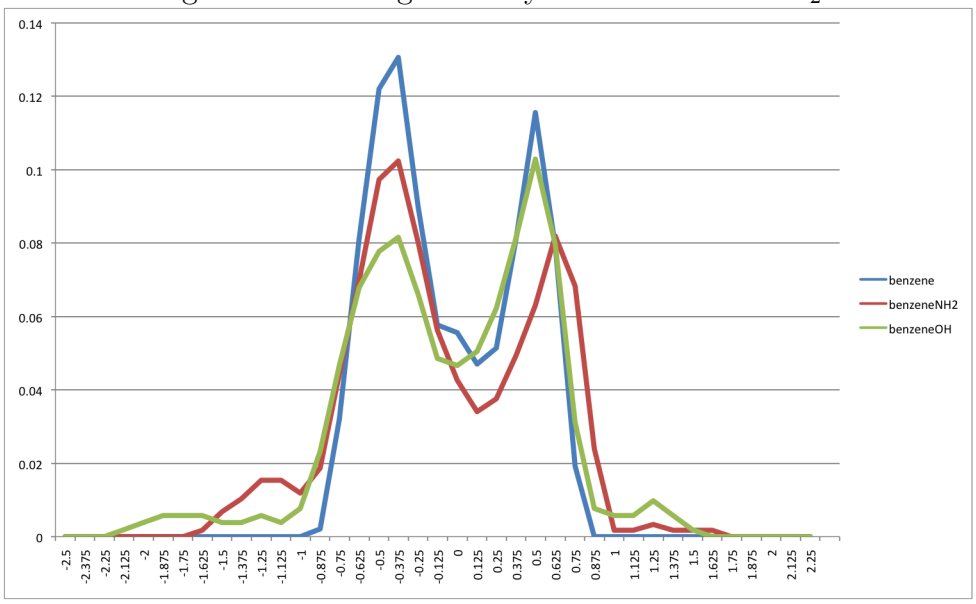


Figure 4.12: Charge Density Profile BenzeneNH_2



donate more electron density in the p system, but this was NOT seen in OH and NH_2, perhaps because the differences in relative surface areas compensated for what can be attributed from the electron donating effect of the substituent.

Category II substituents

Next, an example from category II substituents was chosen, nitrobenzene, for analysis of solvent fingerprint surfaces compared to benzene. Since the resonance structures associated with this class of substituents shows a partial positive charge on the substituent, they have strong electron-withdrawing character. In addition, through the lone pairs, they have resonance contribution

Figure 4.13: Charge Density Profile BenzeneMeO

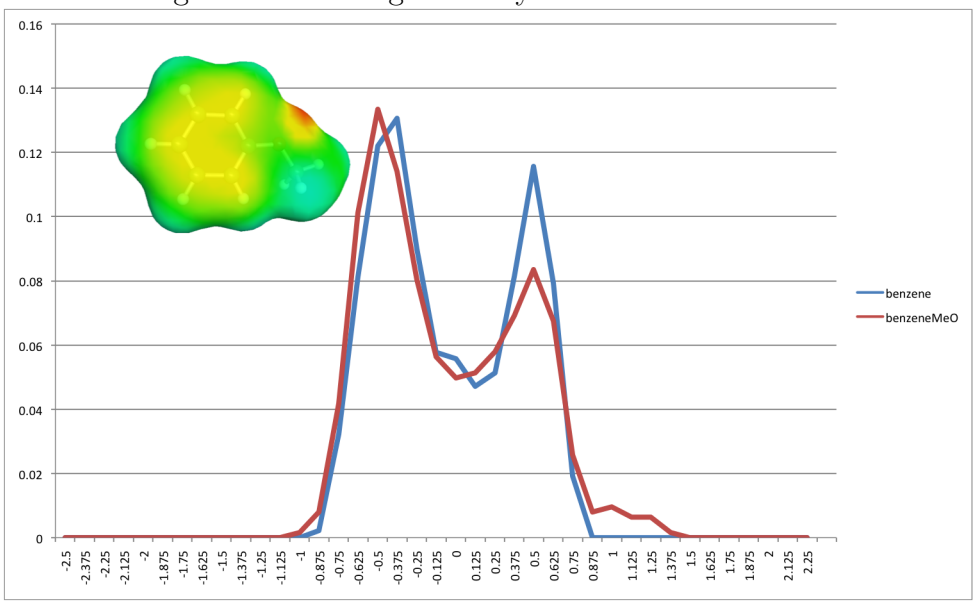
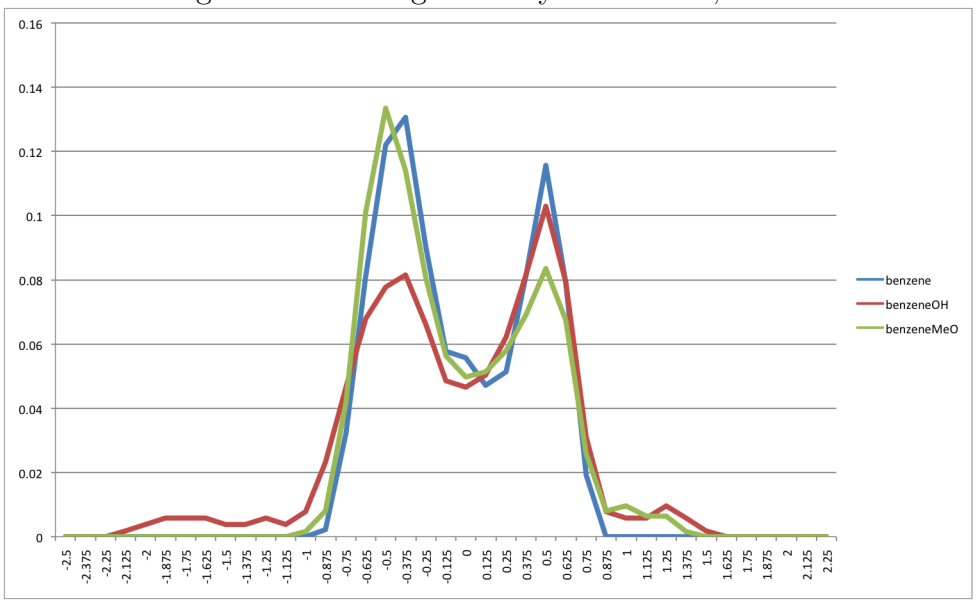


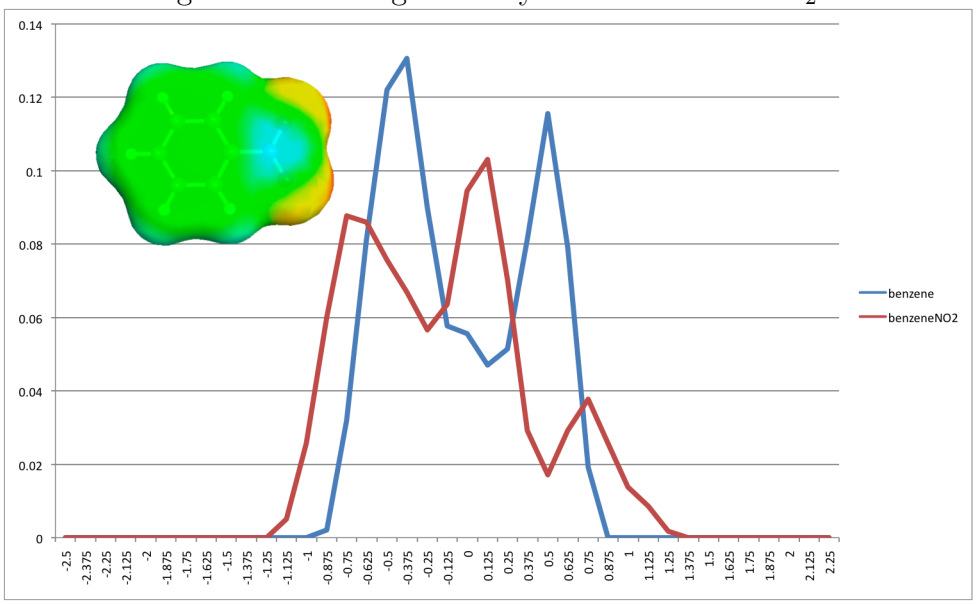
Figure 4.14: Charge Density Profile OH, MeO



that activates the ortho- and para positions of the benzene ring.

The electron withdrawing ability of NO_2 is particularly obvious, as the main peak associated with the pi density is essentially neutralized, showing up as a nonpolar peak set at zero. There also appears a strong peak at 0.875 due to the negative oxygen atoms, which are associated with hot areas for nucleophilic attack. Additionally, the effect of the substiuent decreases the positive area associated with the hydrogen belt, and instead a large contribution from the nitrogen partial positive charge appears as a polar peak associated with nucleophilic behavior. One sees a faint trend towards higher negative values (yellowish) in the meta/para positions of the molecular solvent charge density graphic, and again, compared to benzene, this is a much smaller area of

Figure 4.15: Charge Density Profile BenzeneNO_2



partial negative charge.

Category III substituents

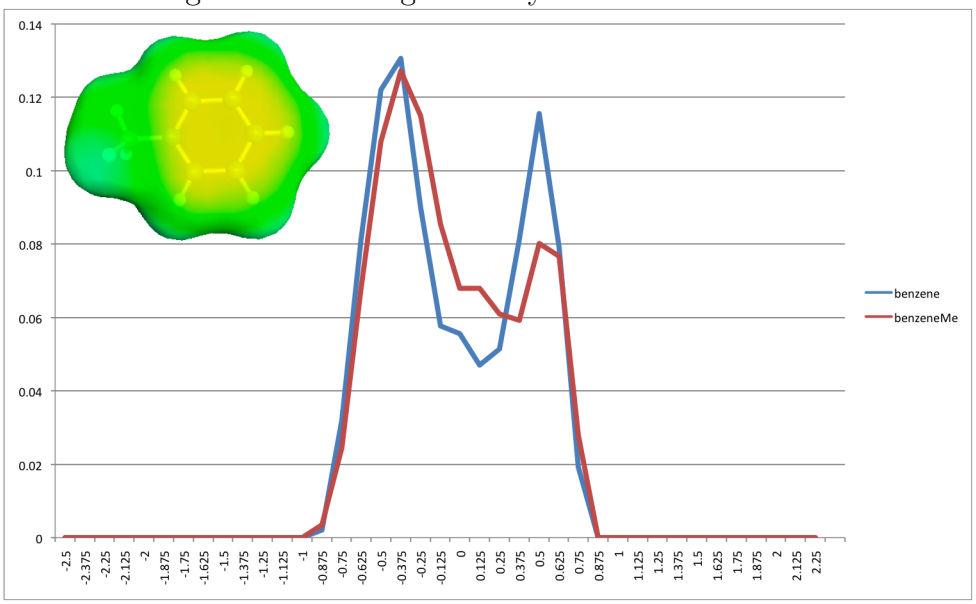
Addition of a methyl substitutent to benzene results in only a mild perturbation to the benzene ring. There is only a slight inductive effect due to the methyl substitutent. Additionally, any resonance effects due to hyperconjugation are very weak. The peak in the electrophillic region is slightly larger and less positive (-0.5nm rather than -0.6nm) compared to benzene, presumably due to the additional hydrogen partial positive charges and the slight inductive effect through the sigma frame, resulting in the charge density of the hydrogens are slightly less polar. The peak in the nucleophilic region is much more focused, albeit smaller in magnitude, and indication that the pi face is less delta positive than in benzene. There is also an additional peak of larger magnitude in the electroneutral area (e.g., around zero).

Category IV substituents

Two category IV substituents were analyzed in terms of their solvent surface fingerprint, CF_3 and $+NH_3$. These substituents have strong electron-withdrawing effects through the sigma system (inductively), either due to the partial charge ($+NH_3$), or due to strong electronegative groups (CF_3). There are no resonance effects with these substituents, since there are no orbitals or electron pairs that can donate into the ring.

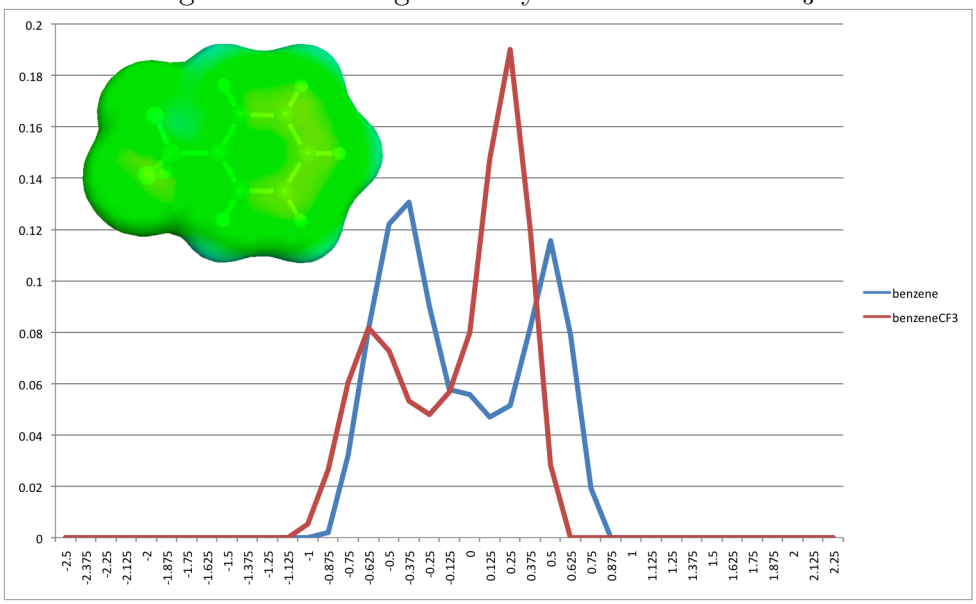
With the CF_3 substituent the peaks shift to the left and are dramatically different in surface area. Like NH_3, the electrophilic peak corresponding to the positively charged H's is shifted to the left (more polar) but with a surface area that is much less. The nucleophilic peak that corresponds to the negatively charged pi face is shifted to the left considerably towards the

Figure 4.16: Charge Density Profile BenzeneMe



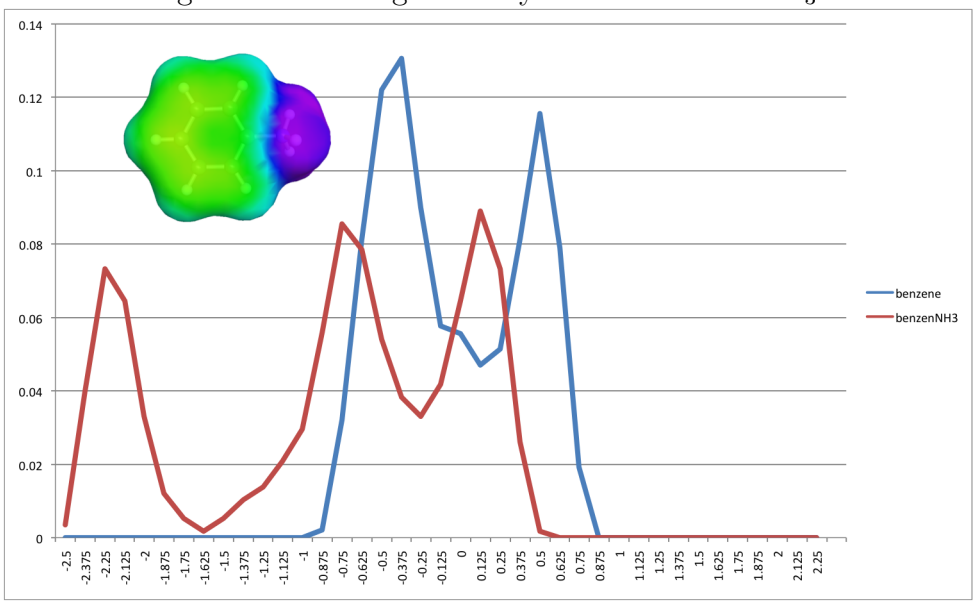
non-polar region, so the charge densities associated with the pi face is more non-polar, but there is a much greater surface charge density associated with the pi face. This effect could be due to the overall increased surface area of the molecule.

Figure 4.17: Charge Density Profile BenzeneCF_3



In the case of +NH3, the resulting substituted benzene is a cation. This strong positive charge is quite evident in the sigma profile, which shows a large peak far to the left, resulting from this positively charged group (also shown as purple with a light blue halo on the graphic). In general, there is significantly more of the surface area of benzene that is associated with positive charge (left region of the graph), including the shift of the pi region towards a more neutral direction, a further indication of the withdrawal effect of this substituent.

Figure 4.18: Charge Density Profile BenzeneNH_3



4.4 Conclusions

For the localized descriptor tests, the density differences at 0.7 Bohr describes the ortho-meta-para directing ability more clearly across the set of mono-substituted benzenes used than than the molecular orbital based descriptors. The patterns were also evident in the normalized and un-normalized nucleophilic and electrophilic frontier orbital densities, but not as clearly for the more weakly activating/deactivating substituents. The disadvantage of using the frontier orbital densities is that this metric is only useful for comparisons across a single molecule so would not be useful for the formulation of a QSAR model based on the analysis of a series of molecules. The electrophilic superdelocalizabilities represented the expected reactivity pattern for the ortho and para directors, but not for the meta directors. For all the MO based descriptors, the nucleophilic analog did not show the patterns as clearly as the electrophilic one.

The solvent charge density fingerprint provides a chemically intuitive way in which to describe the polarity, and thereby the reactivity of the molecule. The expected trends for the various categories of substitutents were illustrated by the representative plots of substituents from these categories, showing the known general trends in polarity.

Chapter 5

QM Interactions in a Host-Guest Artificial Receptor

5.1 Introduction

Intermolecular interactions involving aromatic rings have been shown to be of great importance in biological recognition.[12] The pharmacophore for agonist and antagonist binding in most GPCR receptors show one or more aromatic rings when there is measurable affinity. For example, the antagonist binding site of 5HT2a is lined with aromatic residues that are thought to provide stabilization through the interaction with the aromatic rings in associated ligands (e.g. AMDA, discussed in Ch. 7). Investigations aimed at an energetic quantification of individual interactions with aromatic rings in biological complexes are essential to understanding the mechanism. This also involves an in-depth understanding of all contributing non-bonded interactions such as H-bonding, dispersion, and entropic effects, and how these interactions contribute energetically to binding and recognition. Many studies aim to break down the process of molecular recognition into components by analyzing the quantitative structure activity relationships. However, with many degrees of freedom of these large systems, together with uncertainties in our general knowledge of the details of the biological systems, makes it difficult to break down the interactions quantitatively. Many studies have employed the use of synthetic receptors in an attempt to simplify the problem. In this way, the components and magnitude of individual nonbonded interactions can be probed using high-level computational studies.[10] Often, the high-level analysis of molecular structures can reveal information that was not possible to determine experimentally. For example, Parac et al.[11] analyzed a host-guest system using density functional theory with empirical corrections for dispersion interactions using energy decomposition analysis. Their results indicated that nonspecific van der Waals dispersive interactions were far more important than previously thought. For this reason, an artificial receptor system was chosen to explore interactions important to binding and stabilization in the known complimentary biological model.

5.2 Data-set and Computational Details

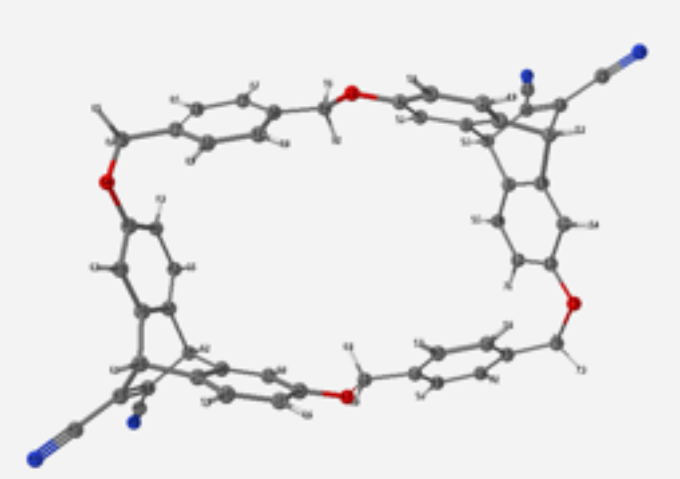
The artificial receptor model is based on a cyclophane host that can "house" small-molecules, called guest molecules. As such, this type of model is often referred to as a "host-guest" model. The original experimental paper that explored such a model, postulated that guest molecules were stabilized through an interaction called the cation-pi interaction.[12] Cation-pi interactions being the stabilizing interactions between a cation and the negatively charged face of a benzene ring. Experimentally, it is difficult to quantitatively break down the various contributions to binding, which include sterics, electrostatics, dispersion, and solvation. The solvation-desolvation phenomena are quite complex and yet very important to account for, since the experiments are carried out in solution.

Aromatic interactions such as polar-π, cation-π and π-π, are thought to be important in molecular recognition in general. The model cyclophane hosts and the guests chosen for this work have been used extensively to try to quantify the various binding interactions involving π systems. Early studies showed that these artificial hosts preferred aromatic guests over aliphatic guests and it was thought that it was the π-interactions that were the driving force.[13] Direct quantitative confirmation of these weak interactions is difficult and confirmation of their importance comes from statistical analysis of their crystal structures which reveal, for instance, that NH bonds tend to point towards the faces of aromatic systems

An objective of the present computational study of a similar host-guest system that have been investigated previously, was to determine more specifically what effects are important in such systems, exploring what energetic factors contribute to the relative binding energetics of the guests. Much of the earlier work focused on only one or two specific aspects of the various interactions, and are therefore somewhat biased in their interpretation. The approach used in this work included a set of molecular systems that enable the examination of the contribution of all factors contributing to the effect to the binding of a guest molecule into a host molecule. High-level quantum chemical computations are used to carry out this detailed investigation, enabling a detailed look into the structure space (steric effects), the molecular electronic space extent (electrostatics), and the effects of solvation (environmental effects).

The cyclophane host scaffold (Figure 5.1) under investigation is one that is composed primarily of aromatic rings. The arrangement of the rings is such that there is a large inner cage. Thus, the host provides a hydrophobic bonding cavity that can accommodate 'guest' molecules of the appropriate size. The dimension of the host is approximately 11.0 by 6.6 Å. The two long sides of the host have two aromatic rings, while the short sides have 1 aromatic ring, with the OH_2 connections in the center of the long side and two opposing corners, and a triptycene-like connection in the remaining opposing corners. When the host is empty in gas phase, B97D/DZV(2d,p) calculations show the long side to be quite twisted, as shown in Figure 5.1.

Figure 5.1: Cyclophane "Host"



Three guest molecules (Figure 5.2) were investigated in this study, quinoline (parent), 4-methylquinoline (neutral), and N-methylquinoline (cation). The neutral and cation guests have approximately the same size, shape, and hydrophobic surface area, but their response in solvent environment is expected to be quite different due to the placement of the CH_3 substituent as well as the resulting molecular charge.

These particular guest models are analogues of guest investigated in previous experimental studies of Dougherty.[14] The modified representations of the original guest were included to specifically investigate the importance of electrostatic, steric, and solvation effects, and their importance in the so-called cation-pi interaction.

The guests are approximately 6.8 Åwide and 3.9-5.8 long. The largest extent refers to the cation and neutral guests with the additional methyl group. Comparing these dimensions to that of the host, one finds no direct steric clashes for the guests to fit into the cavity. However, the polarity of the guest should have important consequences to the complementarity in the host environment, and the preference for the guest to be in water environment vs the more multipolar (e.g., quadropoles of the aromatic components) environment inside the host.

Figure 5.2: quinoline (parent), N-methylquinoline (cation) and , 4-methylquinoline (neutral) guest molecules

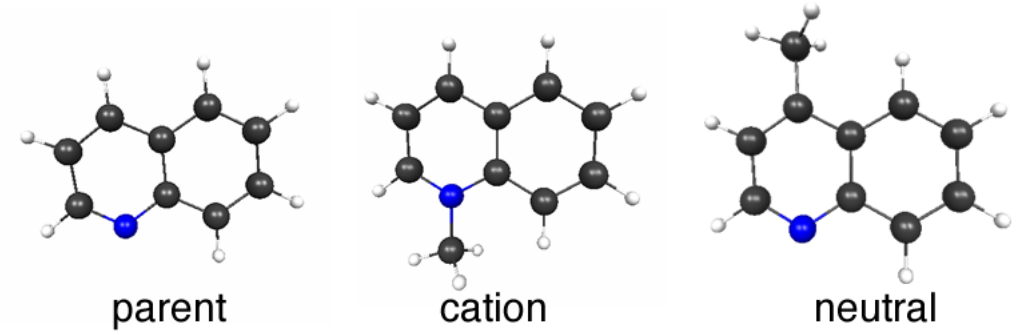


Table 5.1: B97D/DZV(2d,p) complexation energies in gas phase for the full host-guest complexes in kcal/mol

guest	E	E + ZPE
cation	-38.947	-37.412
neutral	-30.537	-29.256
parent	-27.205	–

5.3 Results and Discussion

Geometry optimizations of the various guests in the host were originally done in the gas phase at the RHF/DZV(2d,p)level of theory. During the optimization, the neutral guest molecule slipped out of the host, as shown in Figure 5.4. This behavior was observed for all three guests, but much more pronounced for the neutral (Figure 5.4) and parent molecules (Figure 5.8). The cation molecule only shifted slightly outside the host (Figure 5.6). Alternative starting orientations, with the cation and neutral guests having the methyl group towards the outside of the host cavity, were also investigated. However, even with the reverse orientation, the cation guest nitrogen and methyl group preferred to stay in the host cavity, while the neutral guest preferred to have the polar nitrogen and methyl group outside the host. The parent structure (without a methyl) also preferred to have the polar nitrogen outside the cavity.

Because the RHF level of theory neglects effects of dispersive interactions, it could be possible that the guest molecules move outside of the host cavity due to this missing component in the model. As such, an analogous set of computations were carried out using the MP2/DZV(2d,p) level of theory, a conventional perturbative approach to the inclusion of solvation, as well as the B97-D/DZV(2d,p) level of theory, a dispersion enabled density functional theory approach. However, in both of the latter methods, similar results were found in terms of geometric effects. Geometry optimizations in gas phase all lead to ligands slipping out of the host, and therefore likely result from steric repulsion that can not be pacified in the gas phase environment. Interestingly, gas phase geometry optimization of the cation with starting orientation outside the cavity resulted in an optimized complex having the cation inside the host (Figure 5.10), another indication that there is electrostatic stabilization due to the interaction with that guest. In the case of the neutral guest, geometry optimization resulted in a preferred orientation with the polar nitrogen (electrostatic component) and methyl group (steric component) outside the host rather than inside.

5.3.1 Shape/Sterics, Gas Phase

Table 5.1 lists the calculated gas phase complexation energies for the full host-guest complexes for three guest molecules, in kcal/mol. The complexation energy = $E_{host-guest}$ - (E_{host} + E_{guest}).

Figure 5.3: Host-Neutral Complex: Starting coordinates

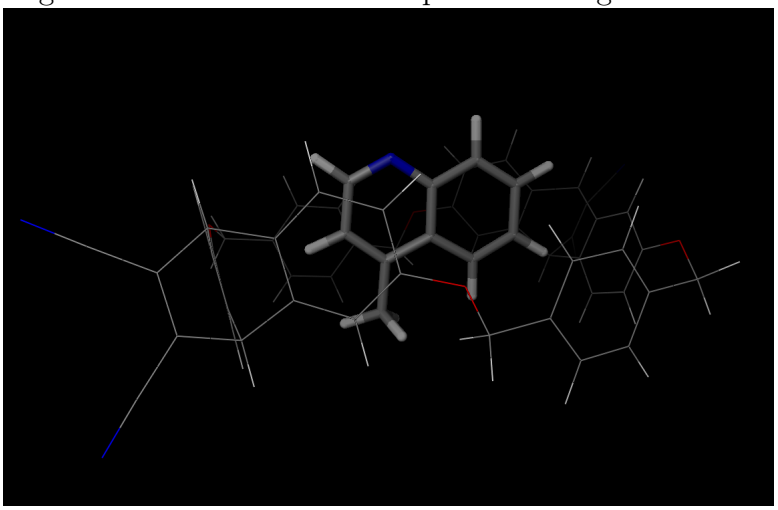


Figure 5.4: Host-Neutral Complex: Optimized coordinates

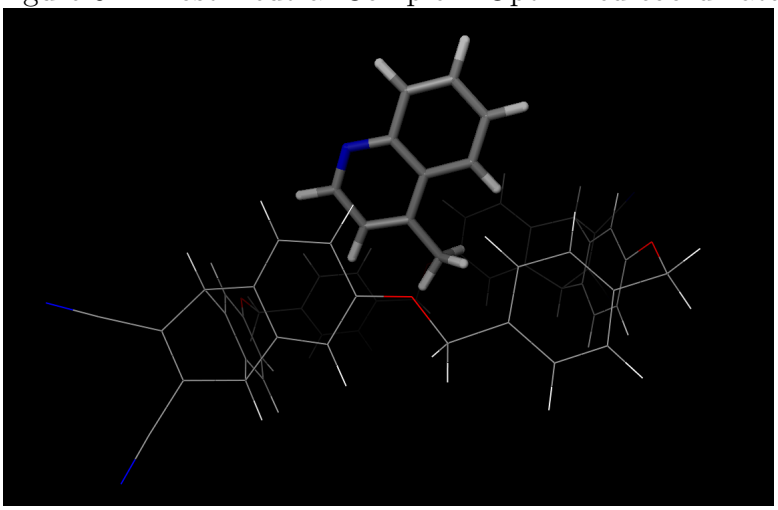


Figure 5.5: Host-Cation Complex: Starting coordinates

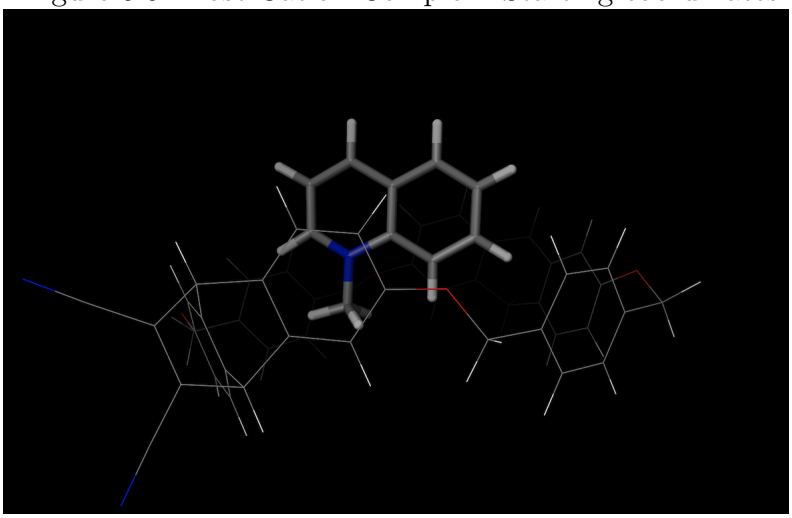


Figure 5.6: Host-Cation Complex: Optimized coordinates

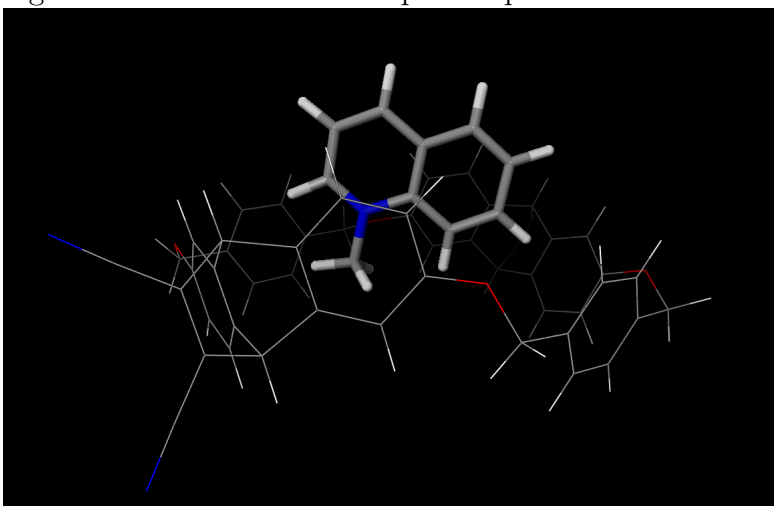


Figure 5.7: Host-Parent Complex: Starting coordinates

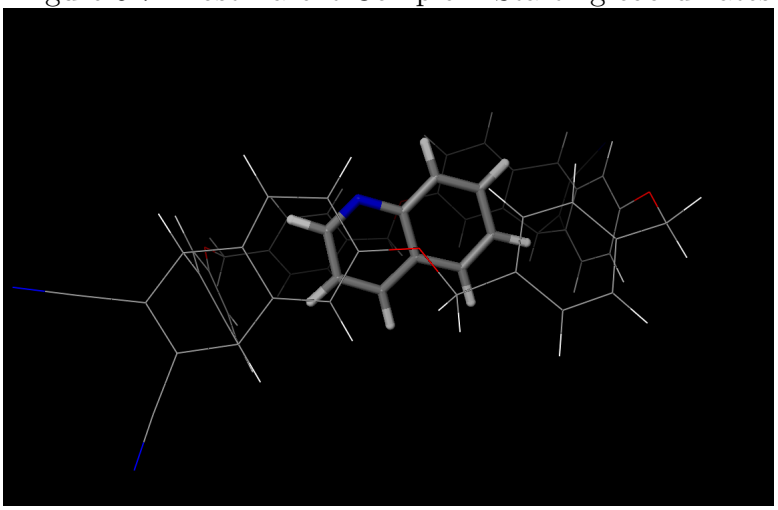


Figure 5.8: Host-Parent Complex: Optimized coordinates

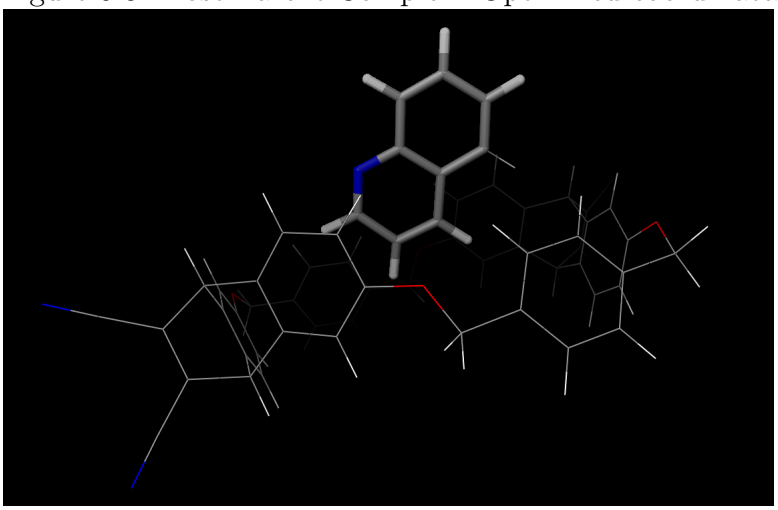


Figure 5.9: Host-Cation-Complex, flipped and inverted: Starting coordinates

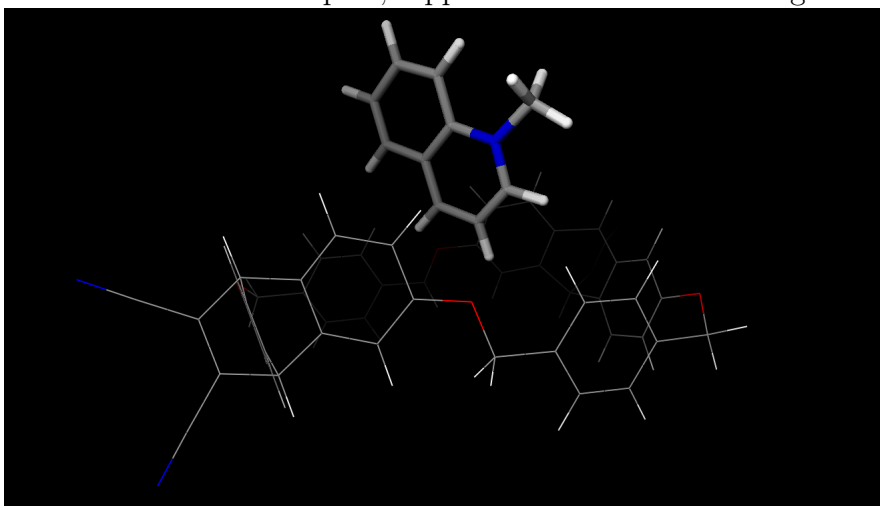
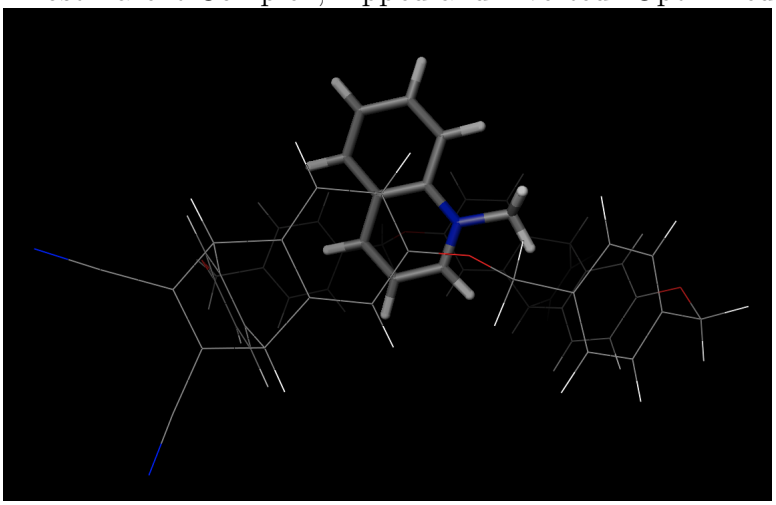


Figure 5.10: Host-Parent Complex, flipped and inverted: Optimized coordinates



5.3.2 Effects of Solvation

In addition to steric and electrostatic factors related to the host-guest interaction, there is an energetic cost associated with the movement of the guest from the solvent into the cavity. Is it more energetically favorable for the guest to be stabilized by solvent, or by the host? What is the cost of breaking the network of interactions of guest with solvent, in order to move into the host?

To help answer these questions, it is instructive to look at the dipole of the various components of the total host-guest system (Table 5.2). The cyclophane host is apolar but the presence of a guest with a permanent dipole will add an attractive dipole-induced dipole interaction.[15] The orientation of dipole can be important in providing steric advantage for slipping into the host. The dipole of the cation and neutral are similar in magnitude at 2.41 D and 2.48 D, respectively, but oriented differently. The dipole of the cation is oriented diagonally with the positive end at the Nitrogen, while the neutral guest has a dipole along the shorter side of the

Table 5.2: Dipole Moments (Debye)

guest	gas phase	solvated
cation	2.41	3.75
neutral	2.48	3.79
parent	1.96	3.07

molecular axis. The parent compound has a dipole of 1.96 D, also oriented along the short molecular axis. The dipole moments of the the three guests is depicted in Figures 5.11, 5.12, 5.13) , where the arrow points in the positive direction.The magnitude of the molecular dipole provides one indication of the expected solvation effects. A second important consideration is charged vs neutral system, where one expects a larger effect of solvation in the former case.

Figure 5.11: B97D/DZV(2d,p) gas phase Cation Dipole

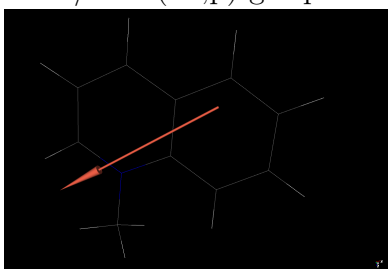


Figure 5.12: B97D/DZV(2d,p) gas phase Neutral Dipole

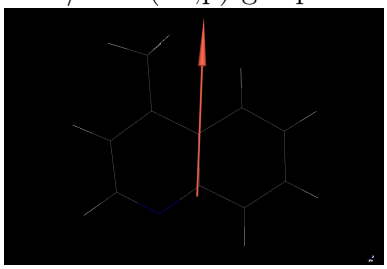
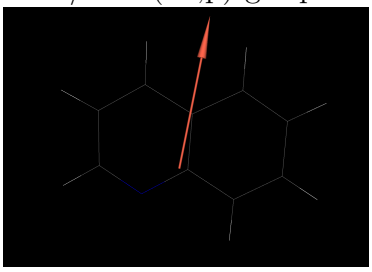


Figure 5.13: B97D/DZV(2d,p) gas phaseParent Dipole



The B97D/DZV(2d,p) calculated results in water environment are shown in Table 5.4. Upon solvation of the guest molecules alone, the orientation of the dipole remains approximately the same in all three, however, the magnitudes increases, as might be expected in solution

Table 5.3: B97D/DZV(2d,p) solvation energies of the three guest molecules, calculated as E_{sol} E_{gp} in (kcal/mol)

molecule	E_{solv}	$(E+ZPE)_{solv}$
cation	-44.32	-44.24
neutral.	-4.65	-4.16
parent	-3.95	-3.90
host	-35.16	-35.49

Table 5.4: B97D/DZV(2d,p) complexation energies in water environment in (kcal/mol)

molecule	E_{solv}	$(E+ZPE)_{solv}$
host-cation	-37.603	-36.064
neutral.	-29.565	-27.968
parent	-26.609	-25.538

environment. In water, the dipoles increases by 1.3 D (cation), 1.3 D (neutral) and 1.1 D (parent). The difference in dipole from gas phase to solution phase is indicative of the expected solvent effects, and is indicative of the preference for the guest to remain in the host or prefer to escape to the water environment.

The stabilization of the three guests due to the solvent environment is calculated as $E_{sol} = E_{sol} - E_{gp}$. As expected, one sees the largest solvation energy associated with the cation guest, a -44.2 kcal/mol, while a much more modest solvation stabilization is found for the neutral and parent guest, -4.16 and -3.90 kcal/mol, respectively.

Interestingly, relative values of complexation energy predictions across the three system does not change significantly from gas phase to solution phase (Table 5.1 vs Table 5.4). The solution phase values are slightly attenuated from their respective gas phase values (e.g., -27.2 vs -26.6; -29.2 vs -28.0; -37.4 vs -36.1, for parent (without ZPE), neutral, and cation guest, respectively).

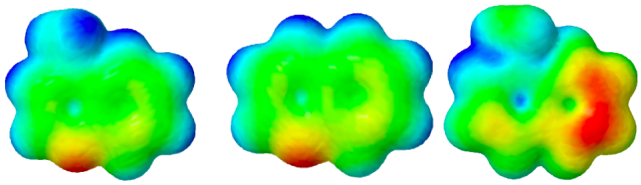
It is interesting to note that, in the solution phase optimization of the host-guest complex, the guest no longer moves from inside the host to outside the host, as observed in the gas phase optimizations. Additionally, during the geometry optimization with the starting position of the cation outside the host, the cation shifted such that the shorter axis entered the host first, corresponding to the orientation of the dipole.

An alternative way to assess the reactivity of a molecule towards positively or negatively charged reactants is by looking at the molecular electrostatic potential, or MEP. The MEP is determined at any given point of the molecules electronic structure and represents the force acting on a positive test charge (e.g., proton) located at that point through the electronic charge cloud of the molecule. No polarization is effected with the test proton, but the resulting electrostatic potential is still a good gauge for assessing the response of a molecule towards a reactant species. Here the MEP for the host and the three guests at the B97D/DZV(2d,p) level

of theory in water environment have been analyszed to investigate the potential complementarity of the host/guest complexation process. For specific charges at atoms, additional calculations were carried out using the CHELPG (CHarges from Electrostatic Potentials using a Grid based) method, which provides a rough picture of the partial charges on individual atoms.

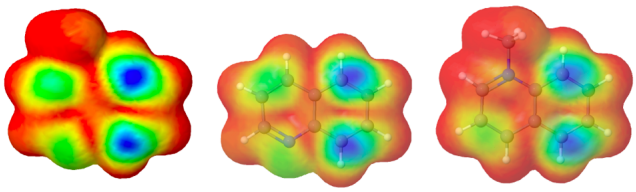
The three guest MEP maps are shown in Figure 5.14. The parent and neutral guest are relatively similar in this perspective, with the latter having the extra electrophilic feature associated with the methyl substituent, but both showing an electron rich site where the nitrogen sits. The positively charged N-methylquinoline is quite different on the other hand, having a charge polarization towards the unsubstituted ring of the quinolone, and in this case with the nitrogen in the position bonded to the methyl group, where the inductive effects of the methyl group result in a slightly more electron rich substituent.

Figure 5.14: B97D/DZV(2d,p) molecular electrostatic potential maps in water environment for 4-methylquinoline, quinoline, and N-methylquinoline cation, guest species.



Focusing on the highest occupied molecular orbital, which would be expected to participate in reactivity, one can look at the electrophilic HOMO density. Such a map shows the highest probability of attack or interaction with an electrophile (dark blue). In this case, one sees a more similar picture of the face of the fused ring system for the three different guest molecules, in addition to the moderately reactive nitrogen site in the parent and neutral system.

Figure 5.15: B97D/DZV(2d,p) molecular electrostatic potential maps in water environment for 4-methylquinoline, quinoline, and N-methylquinoline cation, guest species.



In general, experimentally it is known that the solubility of the cation (0.52 M) is greater than the parent or neutral guests (0.078 M and 0.014 M respectively).[16] If hydrophobicity were the dominant factor behind the binding of the guests to the host, then the more water soluble cation should bind significantly less than the parent or the guest. However, there are also other factors to consider, such as steric and electrostatic effects.

The process of desolvating a guest molecule for entrance into the hydrophobic host will cause an increase in entropy from the disruption of aligned water molecules, increasing the overall G of

binding. In particular, the cation would be considerably more stabilized by solvent than either the neutral or parent guest molecules, which would contribute to the potential for binding of the cation into the host.

The partial charge (Table 5.5) of the Nitrogen for the cation is 0.006 a.u. while the neutral and parent are both negative with the neutral being -0.5934 and the parent being -0.5314. Overall the MEPs of the neutral and the parent species look much more similar than for the cation. Note that the figures below are created using the approximated partial charges from Molekel mapped to the solvent accessible surface while the charges listed below are from the CHELPG algorithm of GAMESS. The trends are approximately the same but the CHELPG charges can be considered more accurate (equations in next section). It is important to pay attention to the color key because the ranges for the color representation is different for each molecule.

Figure 5.16: B97D/DZV(2d,p) calculated CHELPG charges for 4-methylquinoline, quinoline, and N-methylquinoline cation, guest species

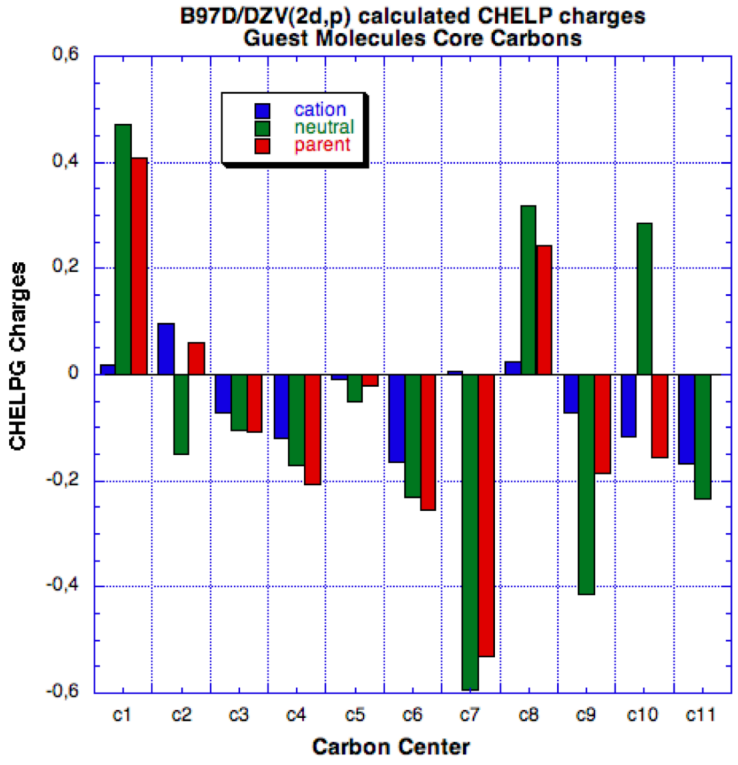


Figure 5.17 shows the MEP map for the isolated host system. From this perspective, one sees that one side of the host is more hydrophobic than the other. The complementarity between the hydrophobic regions of the hosts with each of the guests can be seen in the remaining MEP maps (Figure

Table 5.5: CHELPG partial charges

atom	cation	neutral	parent
c1	0.1753	0.4717	0.4077
c2	0.0967	-0.1484	0.0595
c3	-0.0721	-0.1041	-0.1069
c4	-0.1193	-0.1719	-0.2079
c5	-0.0079	-0.0507	-0.0205
c6	-0.1648	-0.2306	-0.255
n7	0.006	-0.5934	-0.5314
c8	0.0243	0.3172	0.243
c9	-0.0724	-0.4145	-0.1855
c10	-0.1169	0.2853	-0.1558
c11	-0.1679	-0.2345	–

Figure 5.17: B97D/DZV(2d,p) molecular electrostatic potential surface of the host molecule alone

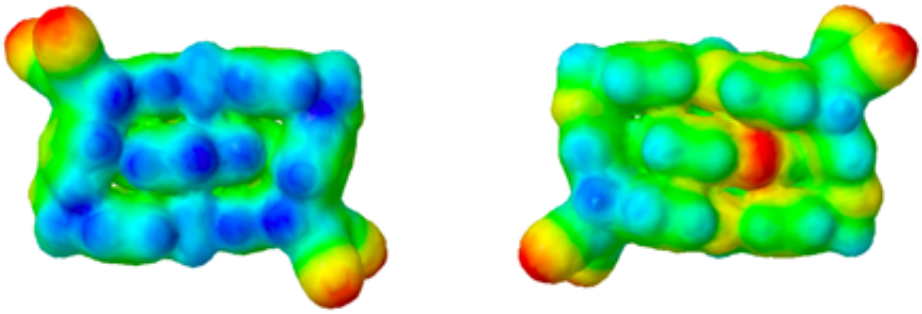


Figure 5.18: B97D/DZV(2d,p) molecular electrostatic potential surface of the host complex with the parent quinolone guest molecule.

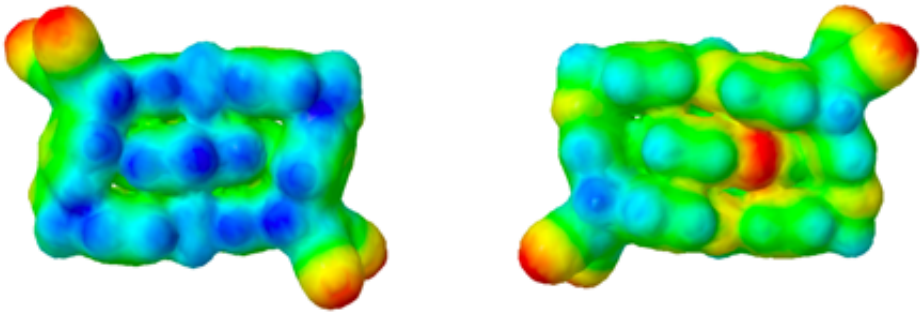


Figure 5.19: B97D/DZV(2d,p) molecular electrostatic potential surface of the host complex with the neutral 4-methylquinoline guest molecule.

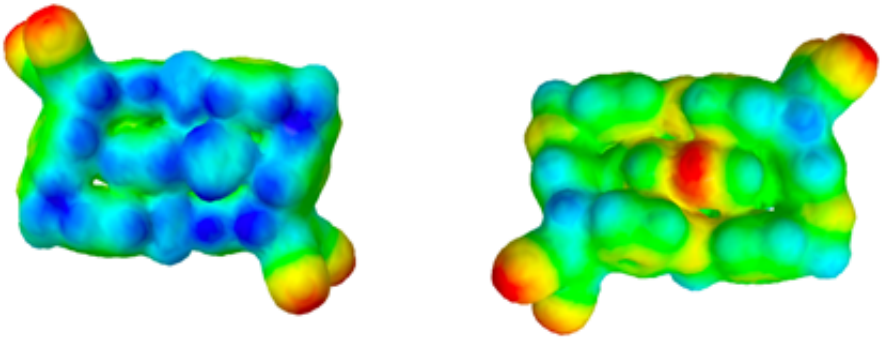
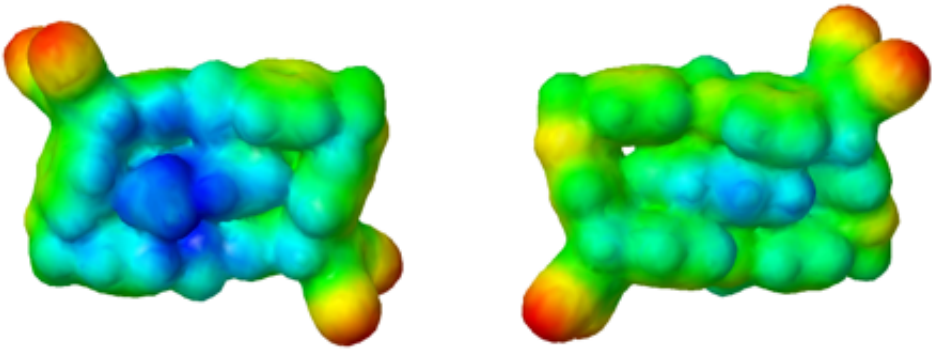


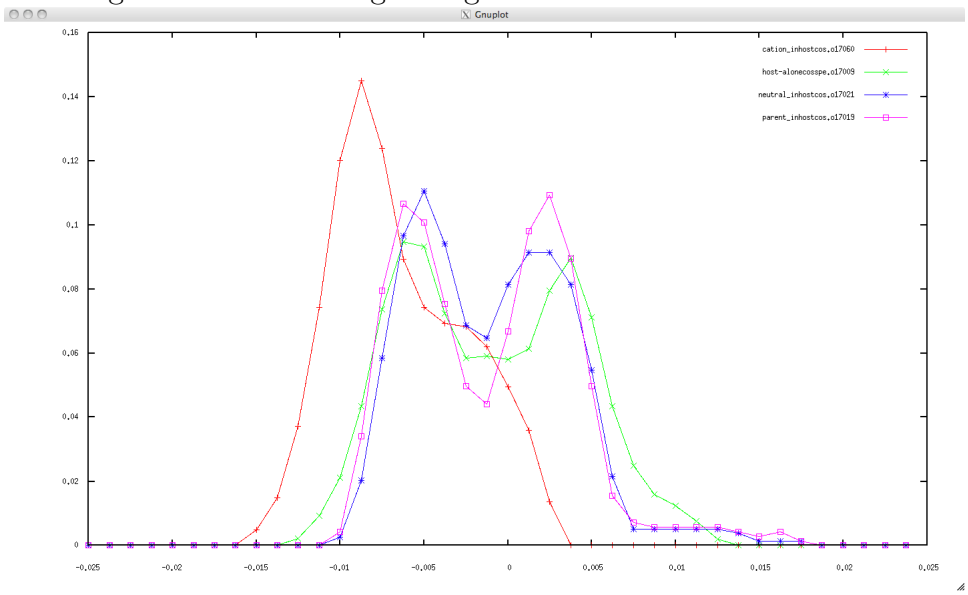
Figure 5.20: B97D/DZV(2d,p) molecular electrostatic potential surface of the host complex with the N-methylquinolinium cation guest molecule



Screening Charge Densities for Host-Guest System

The solvent screening charge density histograms for the host-guest system were evaluated to determine if they could add some insight into the behavior of the host/guest binding. The solvent fingerprint representation of the individual host/guest complex can provide insight into the relative polarity of the guest with respect to that of the host environment. Figure 5.21 shows the histogram plot (sigma plot) of the 3 guests and the host, overlaid in a single graph. This representation shows that the host and the neutral and parent guests share a very similar overall profile. There are two pronounced peaks, one representing nucleophilic behavior in the positive region of the graph, attributed to the side of the guest having the N substituent, and one representing an electrophilic tendency in the negative region of the graph, attributed to the side of the guest having the methyl substituent.

Figure 5.21: Screening Charge Densities of Host and Guests



The cation guest's screening charge density profile has one predominant peak in the positive charge density (electrophilic) region at -0.01 e/nm^2. This peak corresponds to the cation's positive charge. Unlike the host and the other two guests, the cation lacks any prominent peaks in the negative charge density (nucleophilic) region. The host along with the parent and neutral guests all have symmetrically placed peaks at -0.005 e/nm^2 and number of extraneous bumpy features but differ in their relative surface area . Their profiles look quite similar to benzene which should be expected since the host-guest systems are also uncharged aromatic systems. The neutral guest's peak at 0.005 is smaller and broader and there is more of a nonpolar contribution (the valley at 0 e/nm^2 is shallower). The methyl substituent of the neutral guest donates electron density to the ring and causes the negatively charged pi face to become more negative as was the case in methyl substituted benzene. The methyl substituent on benzene also did not have a dramatic effect on the charge density profile.

5.4 Conclusions

The N-methylquinoline (cation) prefers to be in solution rather than in the host ($E_{complexation}$ - E_{sol} = 8.8 kcal/mol) unlike the 4-methylquinoline (neutral) and quinoline (parent) guests which are both more stabilized by the host ($E_{complexation}$ - E_{sol} = -23.81 kcal/mol and -21.64 respectively). This extra stabilization by the host can be attributed to several energetic factors. The hydrophobic effect would stabilize the non-polar neutral and parent guests to a greater degree than for the cation but clearly as the cation binding affinity is larger there must be other factors which compensate for this. Primarily, the significantly larger solvation energy of the cation would contribute favorably to the overall observed binding energy. This energy difference is large enough to account for the difference in binding energy between the cation and the neutral guests and it would not be possible to attribute this to a cation-pi effect specifically.

There are other factors which are thought to stabilize the cation further. Sterically, even though the cation and the neutral are the same size and both can fit into the host, the host must twist and contort significantly to relieve the strain from having the polar Nitrogen inside the cavity for the parent and the guest while not for the cation. This hinted at the possibility of the lone pairs on the nitrogen having a destabilizing effect in the host and that the donor-acceptor interactions are a stabilizing force for the cation and destabilizing for the neutral and parent guests.

Chapter 6

Analysis of Strain in a series of Triptycene Dimers

6.1 Introduction

The standard length of a carbon-carbon single bond is 1.53 Å. Distribution of observed C-C bond lengths from the large set of compounds in the Cambridge structural database shows a very tight distribution centered around this value, and primarily ranging from 1.51-1.55 Å. A well known function, Hookes Law, models bond deformation quite well, $E_b = \frac{1}{2}k(b-b_0)^2$. Where k is the tightness of the bond and dependent on atom type. Close to this 1.53-1.54 Å"normal" deformation of a C-C single bond towards longer bond lengths actually deviates from this perfect harmonic well, however, as weakening the bond does not have such a steep energetic requirement but decreases as the bond lengthens. Such a weakening can occur through electronic effects, through space/through bond effects, or strain. Many examples exist of unusually long C-C bonds.

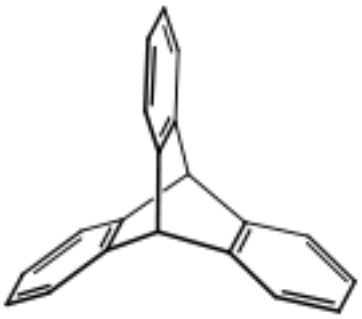
What is less prominent are cases of compressed CC bonds. In conjugated systems, certainly there are "short" representations of the single bonds, due to conjugation of the electrons and rehybridization . However, what about unusually compressed 4-coordinate carbon atoms? One might ask how much can a CC bond be compressed? According to the Hooke's Law model, this should be more energetically costly, particularly since there is not the anharmonic relationship as in the energy bond distances. While small picometer deviations will be relatively easy to accomplish (3-5 kJ/mol), additional compression will require increasingly more energy, e.g. 0.1 Årequires about 35-40 kJ/mol.

In the present investigation, the concept of compressibility of CC bonds was investigated with the idea of better understanding how a common scaffold accommodates compression from a central compressed CC bond.

To make a CC bond shorter, one needs to constrain the carbons together while providing a scaffolding that is sufficiently stiff. The use of axial symmetry can help to focus the compression along the bonding axis. Three-fold symmetric cages[17] are good candidates for investigation

of short C-C bonds. In the present analysis, a series of triptycene dimers, consisting of two triptycenes (Figure 6.1) joined by flanking CC triple bonds was used as the common scaffold. The central atom terminates each monomer subunit with a substituent X and Y. The substituent indicated by the X and Y were systematically varied with all combinations of H, Cl, F, and methyl, and the whole structure analyzed to determine where stress was redistributed as the distance between X and Y increase.

Figure 6.1: Triptycene



6.2 Data Set and Computational Details

Figure 6.2 has the variables of interest on the triptycene scaffold labeled and shows the variant of the standard triptycene scaffold where the central Carbon atom is replaced by a Nitrogen. The n-triptycene variant was used to investigate what effect the electronegative N would have on the compressibility of the CC bonds. The X and Y substituents on the central core of the n-triptycene and the c-triptycene were varied with the 10 combinations of H.F,Cl and Me (Table 6.2) making 20 molecules total. B97D/DZV(2d,p) geometry optimizations of each of the molecules were performed in the gas phase and the relevant bond lengths and angles were parsed from the output using the custom software described in Chapter 8.

A variety of methodology was brought to bear on the assessment of compressibility in the triptycene scaffold as a function of X, Y substitution. The overarching goal was to identify stress synch points in the scaffold. To more effectively evaluate compression uptake, the unstrained system was used as a baseline. In principal, one could choose the monomer, or the dimer with the X=H, Y=H to normalize the compressibility measures. The variable names for the normalized values are the same as the ones listed in Figure 6.2 but with the prefix "dimer_diff" or "monomer_diff" or simply the suffix "_diff" for where there was no corresponding value in the monomer vs. dimer normalization scheme. In this work, the dimer normalization scheme was chosen because there was complete correspondence between all of the length and angle variables which made it easier to compare them statistically. The monomer and monomer differences are discussed in the simple bond analysis, bond distance evaluation however.

The methods employed for the analysis that are described in what follows are broken down into the following types:

1. Simple bond angle, bond distance evaluation
2. Correlation analysis
3. Principle Component Analysis
4. Regression analysis
5. Partial Least Squares
6. Artificial Neural Nets

Figure 6.2: Labeled variables

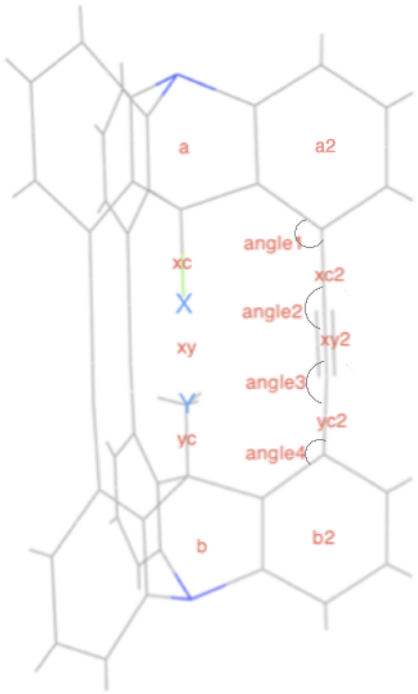


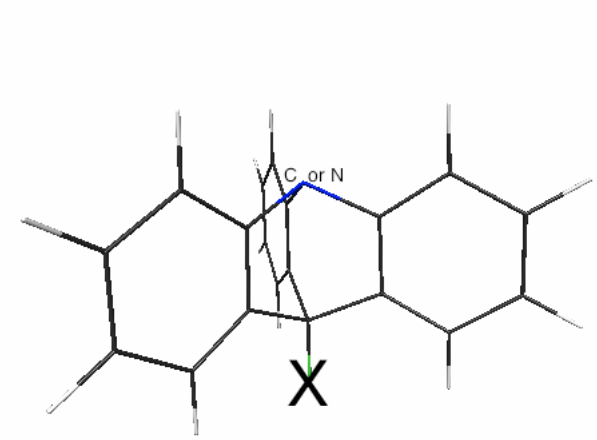
Table 6.1: X,Y Substituents

X	Y
H	H
Cl	Cl
F	F
Me	Me
H	Cl
H	F
H	Me
F	Me
Cl	Me
F	Cl

6.3 Method and Results

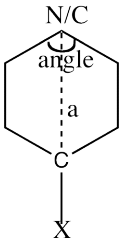
Initial Geometric Assessment

Figure 6.3: Triptycene Monomer



The central ring of the triptycene monomer (Figure 6.3) provides a basis of comparison for what is expected in terms of the length of the bond between the central Carbon atom and the substituent. Table 6.2 lists the lengths of the x-c bond as well as the breadth of the ring and the total length sum (xc + a). The table is ordered according to increasing length of xc. As expected the ranking of the x-c bond lengths are C-H < C-F < C-Me < C-Cl with the N substituted derivative being shorter for each substituent class. The electronegative Nitrogen has a negative inductive effect which shortens the bond lengths slightly. The breadth of the N substituted molecules is a also slightly shorter than the corresponding C-substituted triptycene (except in the case of H) . The central angles are also 0.3-0.4 of a degree more acute for the N-substituted triptycene .

Table 6.2: Monomer Geometry (sorted by XC ascending) in Å

	substituent X	xc	a	Total	angle
	n-H	1.0966	2.6293	3.7258	105.05
	c-H	1.0971	2.6279	3.7250	105.43
	n-F	1.3839	2.6142	3.9981	105.04
	c-F	1.3873	2.6147	4.0020	105.38
	n-Me	1.5395	2.6613	4.2009	105.12
	c-Me	1.5426	2.6644	4.2070	105.52
	n-Cl	1.8010	2.6188	4.4198	105.08
	c-Cl	1.8072	2.6195	4.4266	105.44

In the dimer, the length of xc is affected by the the size and electronegativity of the atom y which is across from it. Table 6.3 lists the respective lengths of xc for each of the atoms

under consideration (where x=H,F,Cl, Me) and ranks the length from smallest to largest. The relatively small F and H atoms have less variation in their xc bond lengths (0.05 and 0.06 Årespectively) while the chlorine atom and methyl group vary by twice as much (almost 0.1 Å). The size of the Y substituent will cause the xc bond length to shorten due to steric repulsion. H-C, for example is the shortest when the H is across from the bulkier substituents methyl and Cl. The same trend is seen for the remaining substituents. Table 6.4 displays the length of xc in the monomer subtracted from the xc value in the dimer to emphasize the differences from the unstrained case. The trend is the same, with the greatest differences in xc length between the strained (dimer) and unstrained (monomer) occurring when the substituent is the bulkier methyl or Cl. Since the strain will be distributed throughout the entire molecule it is insightful to look at what happens to xy and xc+yc as the total length of the central axis increases. Some of the strain will also be absorbed by the flanking ring (Table 6.6). To get a better idea of how the strain is redistributed in the entire system, the variables were sorted according to increasing xc+yc values in Table 6.7 one can visually gauge general trends, the normalized data is shown in Table 6.12. The breadth of the central rings (a and b) contract slightly while the flanking benzene rings (a2 and b2) increase (Figure 6.5), albeit more erratically. The CC single bonds on the flanking axis (xc2 and yc2) range from about 1.43 in the unstrained (X=Y=H) case to about 1.51 (X=Y=Cl) For the Cl and methyl substituents this corresponds to a sizable energetic cost for distorting the bond (in the range of almost a 0.1 Å). The center CC triple bond (xy2) of the flanking ring Å(Figure 6.7) since

The angles of the flanking rings as xc+yc increases are shown in Table 6.8. The angle between the central axis and the benzene rings (angle1 and angle4) increase by approximately 8.6 degrees with the increase in xc+yc (Figure 6.6) while the central bend (angle2 and angle3) does not have a direct correlation to the increase in xc+yc and fluctuate between 174.71(X=Me, Y=H) and 179.99 (X=Y=H). The substituents were ranked according to the normalized (using the respective angles in the unstrained X=Y=H system) average value of this central bend (Table 6.9) to see if there were chemical trends that could explain the data. This central bend angle incurred the most strain (2-4 degrees) when X=Cl or methyl and Y=H. This effect cannot be attributed to sterics alone as the cases where X=Y=Methyl and X=Y=Cl had a very small deviation (> 1 degree) from the unstrained dimer.

The relationship between the distance between the substituents X and Y on the central axis (variable xy) is of particular interest. Table 6.13 ranks the substituents according to increasing values of xy. While the variations in xc bond length in the monomer could be explained using a straightforward argument of sterics and electronegativity, the trend for the x-y length cannot be explained as simply. Figure 6.8 shows the trend in xy as xc+yc increases and the data fluctuates by 0.47 Åas xc+yc increases. One interesting trend is that as the total length of the central axis (total = a + xc+xy+yc+b) increases, the large changes in xc+yc are countered to some extent by xy (Table 6.4) this relationship is easier to see from the plot in Figure 6.4

where the ranking is based on the total length (the actual data for the total is omitted from the graph so that the relationship between the variables xy and xc+yc are clearer). To get a more quantitative measure of how the variables in the system change in response to strain (using xc+yc length as the metric) more extensive statistical analysis was performed and will be described next.

Table 6.3: Ranking X-C length

rank	H-C	F-C	Cl-C	Me-C
1	1.0516 (n-Cl-H)	1.3007 (n-F-Cl)	1.5967 (n-Cl-Cl)	1.3891 (n-Cl-Me)
2	1.0576 (n-Me-H)	1.3039 (n-Me-F)	1.6049 (c-Cl-Cl)	1.3923 (n-Me-Me)
3	1.0576 (c-Cl-H)	1.3076 (c-Cl-F)	1.6057 (n-Cl-Me)	1.3965 (c-Cl-Me)
4	1.0634 (c-Me-H)	1.3110 (c-Me-F)	1.6142 (c-Cl-Me)	1.3995 (c-Me-Me)
5	1.0772 (n-F-H)	1.3154 (n-F-F)	1.6200 (n-Cl-F)	1.4174 (n-Me-F)
6	1.0807 (c-F-H)	1.3232 (c-F-F)	1.6292 (c-Cl-F)	1.4251 (c-Me-F)
7	1.0919 (n-H-H)	1.3572 (n-F-H)	1.6876 (n-Cl-H)	1.4721 (n-Me-H)
8	1.0933 (c-H-H)	1.3602 (c-F-H)	1.6966 (c-Cl-H)	1.4791 (c-Me-H)
range	0.0416	0.0595	0.0999	0.0900

Table 6.4: Ranking X-C length (modulus monomer)

xc	H-C	F-C	Cl-C	Me-C
1	0.0039 (c-H-H)	0.0268 (n-F-H)	0.1105 (c-Cl-H)	0.0635 (n-Cl-Me)
2	0.0047 (n-H-H)	0.0270 (c-F-H)	0.1134 (n-Cl-H)	0.0675 (n-Me-Me)
3	0.0164 (c-F-H)	0.0607 (n-F-F)	0.1780 (c-Cl-F)	0.1175 (c-Cl-Me)
4	0.0199 (n-F-H)	0.0641 (c-F-F)	0.1810 (n-Cl-F)	0.1221 (c-Me-Me)
5	0.0332 (c-Me-H)	0.0762 (c-Me-F)	0.1929 (c-Cl-Me)	0.1431 (n-Me-F)
6	0.0389 (c-Cl-H)	0.0796 (c-Cl-F)	0.1953 (n-Cl-Me)	0.1461 (c-Me-F)
7	0.0389 (n-Me-H)	0.0801 (n-Me-F)	0.2023 (c-Cl-Cl)	0.1473 (n-Me-H)
8	0.0449 (n-Cl-H)	0.0832 (n-F-Cl)	0.2043 (n-Cl-Cl)	0.1505 (c-Me-H)
range	0.0411	0.0564	0.0938	0.0869

Table 6.5: Central Bond Dimer Geometry

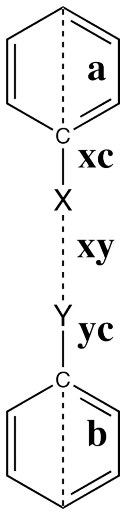
	subst	a	xc	xy	yc	xc+yc	b	total
	n-Me-Me	2.4566	1.3923	2.1637	1.3923	2.7845	2.4566	9.8614
	n-Cl-Cl	2.3917	1.5967	2.0082	1.5967	3.1934	2.3917	9.9850
	n-F-F	2.5338	1.3154	1.9359	1.3154	2.6307	2.5338	9.6341
	n-H-H	2.6219	1.0919	2.0108	1.0919	2.1837	2.6219	9.4383
	n-F-H	2.5822	1.3572	1.8296	1.0772	2.4344	2.6100	9.4562
	n-Cl-H	2.5270	1.6876	1.7184	1.0516	2.7393	2.5346	9.5192
	n-Cl-F	2.4669	1.6200	1.9445	1.3007	2.9207	2.4528	9.7850
	n-Me-F	2.5062	1.4174	2.0010	1.3039	2.7213	2.4865	9.7150
	n-Cl-Me	2.4262	1.6057	2.0455	1.3891	2.9948	2.4407	9.9071
	n-Me-H	2.5778	1.4721	1.7982	1.0576	2.5297	2.5704	9.4761
	c-Me-Me	2.4646	1.3995	2.1917	1.3996	2.7991	2.4648	9.9203
	c-Cl-Cl	2.4059	1.6049	2.0256	1.6049	3.2097	2.4059	10.0472
	c-F-F	2.5471	1.3232	1.9764	1.3232	2.6464	2.5471	9.7169
	c-H-H	2.6225	1.0933	2.1300	1.0933	2.1865	2.6225	9.5615
	c-F-H	2.5889	1.3602	1.9254	1.0807	2.4409	2.6137	9.5688
	c-Cl-H	2.5417	1.6966	1.7625	1.0576	2.7543	2.5501	9.6086
	c-Cl-F	2.4799	1.6292	1.9692	1.3076	2.9368	2.4707	9.8567
	c-Me-F	2.5176	1.4251	2.0334	1.3110	2.7362	2.5015	9.7887
	c-Cl-Me	2.4382	1.6142	2.0686	1.3965	3.0107	2.4491	9.9667
	c-Me-H	2.5882	1.4791	1.8541	1.0634	2.5425	2.5818	9.5666

Figure 6.4: Relationship between xc + yc and xy

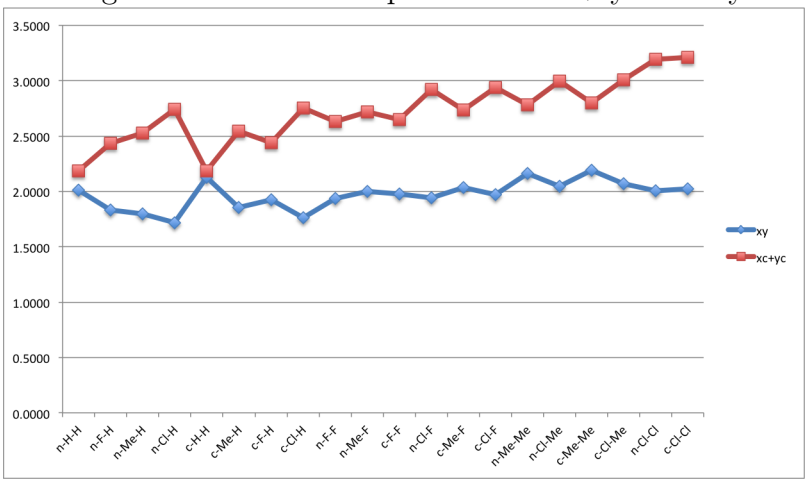


Table 6.6: Flanking Bond Dimer Geometry

subst	a2	xc2	xy2	yc2	xc2 + xy2 + yc2	b2	total2	angle1	angle2	angle3	angle4
n-Me-Me	2.9434	1.4893	1.2479	1.4893	4.2239	2.9433	10.1106	127.6727	177.6102	177.6084	127.6705
n-Cl-Cl	2.9698	1.5099	1.2563	1.5099	4.2748	2.9698	10.2141	128.7965	178.2807	178.2807	128.7965
n-F-F	2.9033	1.4484	1.2331	1.4484	4.1288	2.9033	9.9347	124.2282	178.4410	178.4410	124.2282
n-H-H	2.8592	1.4264	1.2268	1.4264	4.0785	2.8592	9.7926	120.5499	178.4666	178.4666	120.5499
n-F-H	2.8750	1.4314	1.2272	1.4303	4.0860	2.8649	9.8213	122.4651	177.3257	177.5837	121.4735
n-Cl-H	2.8995	1.4519	1.2333	1.4468	4.1217	2.8918	9.9078	125.3692	175.2428	175.0853	124.7366
n-Cl-F	2.9364	1.4825	1.2434	1.4710	4.1936	2.9389	10.0687	126.8974	178.5286	176.1473	127.6105
n-Me-F	2.9186	1.4695	1.2392	1.4618	4.1670	2.9245	10.0097	126.1760	176.0725	178.4695	126.3177
n-Cl-Me	2.9554	1.4991	1.2517	1.4952	4.2435	2.9512	10.1501	127.6960	179.8559	175.8186	128.2624
n-Me-H	2.8818	1.4437	1.2307	1.4396	4.1058	2.8793	9.8615	124.1028	174.7089	176.6840	123.4167
c-Me-Me	2.9237	1.4841	1.2461	1.4841	4.2127	2.9237	10.0574	128.1056	178.1266	178.1179	128.1006
c-Cl-Cl	2.9488	1.5040	1.2546	1.5040	4.2618	2.9488	10.1551	129.1541	178.7967	178.7967	129.1541
c-F-F	2.8859	1.4448	1.2316	1.4448	4.1209	2.8859	9.8919	124.6109	179.2977	179.2949	124.6096
c-H-H	2.8485	1.4262	1.2267	1.4262	4.0791	2.8485	9.7761	121.0488	179.9858	179.9858	121.0488
c-H-F	2.8629	1.4304	1.2267	1.4290	4.0854	2.8526	9.8008	123.0183	178.5182	178.9257	121.9356
c-Cl-H	2.8833	1.4491	1.2321	1.4435	4.1175	2.8752	9.8756	125.8162	176.0561	175.8133	125.1474
c-Cl-F	2.9168	1.4776	1.2418	1.4661	4.1835	2.9190	10.0179	127.2446	179.1892	176.8158	127.9480
c-Me -F	2.9004	1.4654	1.2377	1.4575	4.1584	2.9057	9.9635	126.6199	176.5021	179.3484	126.7015
c-Cl-Me	2.9347	1.4935	1.2500	1.4898	4.2316	2.9310	10.0941	128.0507	179.2425	176.2540	128.6632
c-Me-H	2.8668	1.4415	1.2297	1.4369	4.1025	2.8640	9.8329	124.6321	175.2591	177.7289	123.8727

Table 6.7: Original Variables Ranked According to Increasing xc+yc

subst	xc+yc	a	xy	b	total	a2	xc2	xy2	yc2	xc2+x2y+yc2	b2	total2
n-H-H	2.184	2.622	2.011	2.622	9.438	2.859	1.426	1.227	1.426	4.079	2.859	9.793
c-H-H	2.187	2.623	2.130	2.623	9.561	2.849	1.426	1.227	1.426	4.079	2.849	9.776
n-F-H	2.434	2.582	1.830	2.610	9.456	2.875	1.431	1.227	1.430	4.086	2.865	9.821
c-F-H	2.441	2.589	1.925	2.614	9.569	2.863	1.430	1.227	1.429	4.085	2.853	9.801
n-Me-H	2.530	2.578	1.798	2.570	9.476	2.882	1.444	1.231	1.440	4.106	2.879	9.861
c-Me-H	2.542	2.588	1.854	2.582	9.567	2.867	1.441	1.230	1.437	4.102	2.864	9.833
n-F-F	2.631	2.534	1.936	2.534	9.634	2.903	1.448	1.233	1.448	4.129	2.903	9.935
c-F-F	2.646	2.547	1.976	2.547	9.717	2.886	1.445	1.232	1.445	4.121	2.886	9.892
n-Me-F	2.721	2.506	2.001	2.487	9.715	2.919	1.469	1.239	1.462	4.167	2.925	10.010
c-Me-F	2.736	2.518	2.033	2.501	9.789	2.900	1.465	1.238	1.458	4.158	2.906	9.964
n-Cl-H	2.739	2.527	1.718	2.535	9.519	2.900	1.452	1.233	1.447	4.122	2.892	9.908
c-Cl-H	2.754	2.542	1.763	2.550	9.609	2.883	1.449	1.232	1.444	4.117	2.875	9.876
n-Me-Me	2.785	2.457	2.164	2.457	9.861	2.943	1.489	1.248	1.489	4.224	2.943	10.111
c-Me-Me	2.799	2.465	2.192	2.465	9.920	2.924	1.484	1.246	1.484	4.213	2.924	10.057
n-Cl-F	2.921	2.467	1.944	2.453	9.785	2.936	1.482	1.243	1.471	4.194	2.939	10.069
c-Cl-F	2.937	2.480	1.969	2.471	9.857	2.917	1.478	1.242	1.466	4.184	2.919	10.018
n-Cl-Me	2.995	2.426	2.045	2.441	9.907	2.955	1.499	1.252	1.495	4.244	2.951	10.150
c-Cl-Me	3.011	2.438	2.069	2.449	9.967	2.935	1.493	1.250	1.490	4.232	2.931	10.094
n-Cl-Cl	3.193	2.392	2.008	2.392	9.985	2.970	1.510	1.256	1.510	4.275	2.970	10.214
c-Cl-Cl	3.210	2.406	2.026	2.406	10.047	2.949	1.504	1.255	1.504	4.262	2.949	10.155
range	1.0260	-0.231	0.473	-0.231	0.609	0.121	0.084	0.030	0.084	0.196	0.121	0.438

Table 6.8: Angles of flanking ring sorted according to increasing xc+yc

subst	xc+yc	angle1	angle2	angle3	angle4
n-H-H	2.184	120.550	178.467	178.467	120.550
c-H-H	2.187	121.049	179.986	179.986	121.049
n-F-H	2.434	122.465	177.326	177.584	121.473
c-F-H	2.441	123.018	178.518	178.926	121.936
n-Me-H	2.530	124.103	174.709	176.684	123.417
c-Me-H	2.542	124.632	175.259	177.729	123.873
n-F-F	2.631	124.228	178.441	178.441	124.228
c-F-F	2.646	124.611	179.298	179.295	124.610
n-Me-F	2.721	126.176	176.072	178.470	126.318
c-Me-F	2.736	126.620	176.502	179.348	126.701
n-Cl-H	2.739	125.369	175.243	175.085	124.737
c-Cl-H	2.754	125.816	176.056	175.813	125.147
n-Me-Me	2.785	127.673	177.610	177.608	127.671
c-Me-Me	2.799	128.106	178.127	178.118	128.101
n-Cl-F	2.921	126.897	178.529	176.147	127.611
c-Cl-F	2.937	127.245	179.189	176.816	127.948
n-Cl-Me	2.995	127.696	179.856	175.819	128.262
c-Cl-Me	3.011	128.051	179.243	176.254	128.663
n-Cl-Cl	3.193	128.796	178.281	178.281	128.796
c-Cl-Cl	3.210	129.154	178.797	178.797	129.154

Figure 6.5: Relationship between increasing xc+yc and the ring breadth

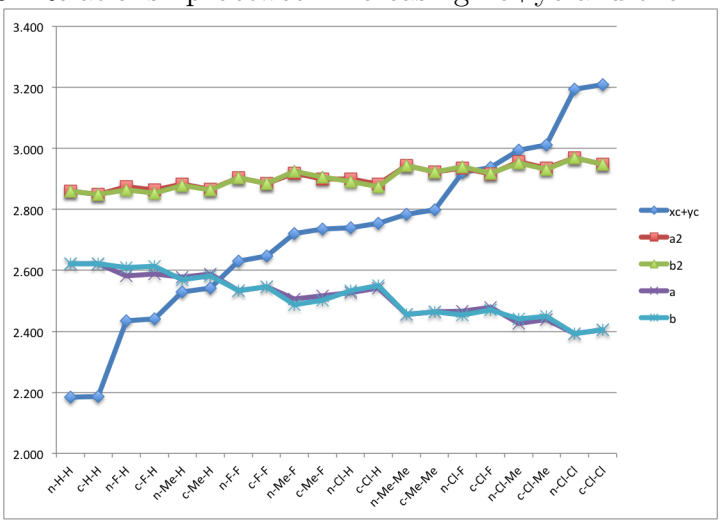


Table 6.9: Substituents sorted according to normalized average value of angle2 and angle3

subst	angle2_diff	angle3_diff	average
c-Cl-H	-3.930	-4.172	-4.051
c-Me-H	-4.727	-2.257	-3.492
n-Cl-H	-3.224	-3.381	-3.303
n-Me-H	-3.758	-1.783	-2.770
c-Cl-Me	-0.743	-3.732	-2.238
c-Me-F	-3.484	-0.637	-2.061
c-Cl-F	-0.797	-3.170	-1.983
c-Me-Me	-1.859	-1.868	-1.864
c-F-H	-1.468	-1.060	-1.264
n-Me-F	-2.394	0.003	-1.196
c-Cl-Cl	-1.189	-1.189	-1.189
n-Cl-F	0.062	-2.319	-1.129
n-F-H	-1.141	-0.883	-1.012
n-Me-Me	-0.856	-0.858	-0.857
c-F-F	-0.688	-0.691	-0.689
n-Cl-Me	1.389	-2.648	-0.629
n-Cl-Cl	-0.186	-0.186	-0.186
n-F-F	-0.026	-0.026	-0.026
n-H-H	0.000	0.000	0.000
c-H-H	0.000	0.000	0.000

Figure 6.6: Relationship between increasing xc+yc and the angles on the flanking ring

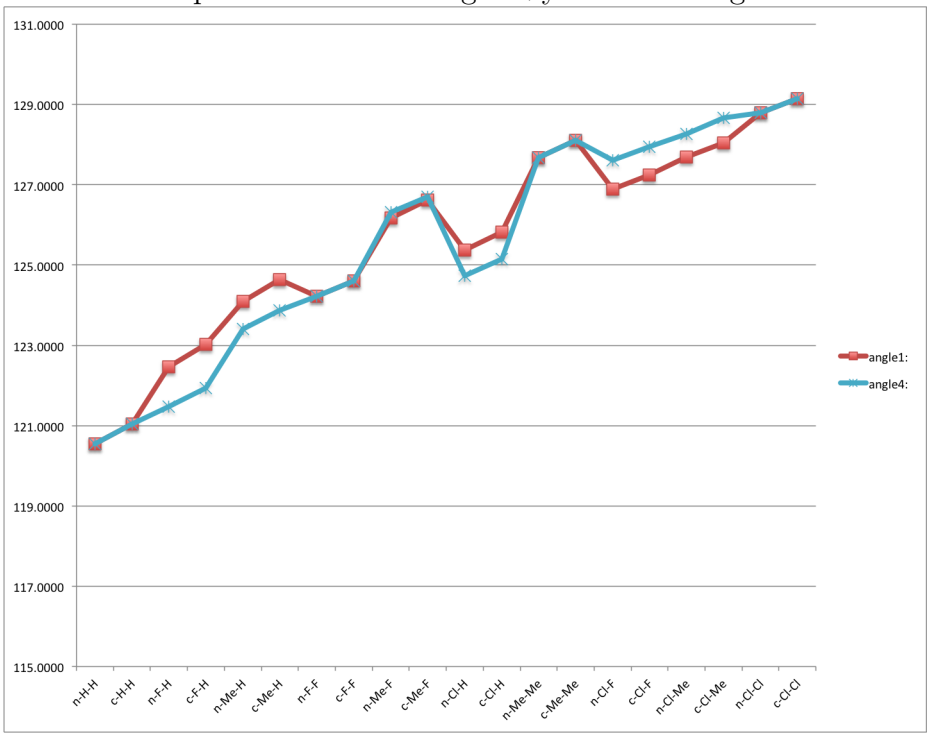


Table 6.10: Normalized Variables (central axis) Ranked According to Increasing xc+yc

subst	xc+yc	a_diff	dimer_diff_xc	dimer_diff_xy	dimer_diff_yc	b_diff	total_diff
n-H-H	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
c-H-H	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
n-F-H	0.2507	-0.0397	0.2653	-0.1812	-0.0146	-0.0119	0.0179
c-F-H	0.2544	-0.0336	0.2670	-0.2046	-0.0126	-0.0088	0.0074
n-Me-H	0.3460	-0.0441	0.3802	-0.2125	-0.0343	-0.0515	0.0378
c-Me-H	0.3560	-0.0343	0.3858	-0.2759	-0.0299	-0.0407	0.0051
n-F-F	0.4470	-0.0881	0.2235	-0.0749	0.2235	-0.0881	0.1959
c-F-F	0.4599	-0.0755	0.2299	-0.1536	0.2300	-0.0755	0.1554
n-Me-F	0.5375	-0.1157	0.3255	-0.0098	0.2120	-0.1354	0.2767
c-Me-F	0.5497	-0.1049	0.3319	-0.0965	0.2178	-0.1210	0.2272
n-Cl-H	0.5555	-0.0949	0.5958	-0.2923	-0.0402	-0.0873	0.0809
c-Cl-H	0.5678	-0.0808	0.6034	-0.3674	-0.0356	-0.0724	0.0471
n-Me-Me	0.6008	-0.1653	0.3004	0.1529	0.3004	-0.1653	0.4231
c-Me-Me	0.6126	-0.1579	0.3063	0.0618	0.3064	-0.1577	0.3588
n-Cl-F	0.7370	-0.1550	0.5282	-0.0663	0.2089	-0.1691	0.3467
c-Cl-F	0.7503	-0.1426	0.5360	-0.1607	0.2144	-0.1518	0.2952
n-Cl-Me	0.8110	-0.1957	0.5138	0.0347	0.2972	-0.1812	0.4688
c-Cl-Me	0.8242	-0.1843	0.5210	-0.0613	0.3032	-0.1734	0.4052
n-Cl-Cl	1.0097	-0.2302	0.5049	-0.0026	0.5049	-0.2302	0.5467
c-Cl-Cl	1.0232	-0.2166	0.5116	-0.1043	0.5116	-0.2166	0.4857

Table 6.11: Normalized Variables (flanking axis) Ranked According to Increasing xc+yc

subst	xc+yc	a2_diff	dimer_diff_xc2	dimer_diff_xy2	dimer_diff_yc2	c2_diff	b2_diff	total2_diff
n-H-H	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
c-H-H	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
n-F-H	0.2507	0.0158	0.0050	0.0004	0.0039	0.0075	0.0057	0.0287
c-F-H	0.2544	0.0144	0.0042	0.0000	0.0028	0.0063	0.0041	0.0247
n-Me-H	0.3460	0.0226	0.0174	0.0039	0.0132	0.0273	0.0201	0.0688
c-Me-H	0.3560	0.0183	0.0153	0.0030	0.0107	0.0234	0.0155	0.0568
n-F-F	0.4470	0.0441	0.0221	0.0062	0.0221	0.0503	0.0441	0.1421
c-F-F	0.4599	0.0374	0.0186	0.0049	0.0186	0.0419	0.0374	0.1159
n-Me-F	0.5375	0.0594	0.0431	0.0124	0.0355	0.0885	0.0653	0.2170
c-Me-F	0.5497	0.0519	0.0392	0.0110	0.0313	0.0793	0.0572	0.1874
n-Cl-H	0.5555	0.0403	0.0256	0.0064	0.0205	0.0432	0.0326	0.1152
c-Cl-H	0.5678	0.0348	0.0229	0.0054	0.0174	0.0384	0.0267	0.0996
n-Me-Me	0.6008	0.0842	0.0630	0.0211	0.0630	0.1454	0.0842	0.3180
c-Me-Me	0.6126	0.0752	0.0579	0.0194	0.0579	0.1336	0.0752	0.2814
n-Cl-F	0.7370	0.0772	0.0561	0.0166	0.0447	0.1151	0.0798	0.2761
c-Cl-F	0.7503	0.0683	0.0515	0.0150	0.0399	0.1045	0.0705	0.2418
n-Cl-Me	0.8110	0.0962	0.0727	0.0249	0.0689	0.1650	0.0920	0.3575
c-Cl-Me	0.8242	0.0862	0.0673	0.0232	0.0637	0.1525	0.0824	0.3180
n-Cl-Cl	1.0097	0.1106	0.0836	0.0295	0.0836	0.1963	0.1106	0.4215
c-Cl-Cl	1.0232	0.1003	0.0778	0.0278	0.0778	0.1828	0.1003	0.3791

Table 6.12: Normalized Variables (angles on flanking axis) Ranked According to Increasing xc+yc

subst	xc+yc	angle1_diff	angle2_diff	angle3_diff	angle4_diff
n-H-H	0.0000	0.0000	0.0000	0.0000	0.0000
c-H-H	0.0000	0.0000	0.0000	0.0000	0.0000
n-F-H	0.2507	1.9152	-1.1409	-0.8829	0.9236
c-F-H	0.2544	1.9695	-1.4676	-1.0601	0.8868
n-Me-H	0.3460	3.5529	-3.7577	-1.7827	2.8668
c-Me-H	0.3560	3.5833	-4.7266	-2.2569	2.8239
n-F-F	0.4470	3.6783	-0.0256	-0.0256	3.6783
c-F-F	0.4599	3.5622	-0.6880	-0.6908	3.5608
n-Me-F	0.5375	5.6261	-2.3941	0.0029	5.7678
c-Me-F	0.5497	5.5711	-3.4837	-0.6374	5.6527
n-Cl-H	0.5555	4.8193	-3.2238	-3.3813	4.1867
c-Cl-H	0.5678	4.7674	-3.9297	-4.1724	4.0986
n-Me-Me	0.6008	7.1228	-0.8564	-0.8582	7.1206
c-Me-Me	0.6126	7.0568	-1.8592	-1.8679	7.0518
n-Cl-F	0.7370	6.3474	0.0620	-2.3193	7.0606
c-Cl-F	0.7503	6.1958	-0.7966	-3.1700	6.8992
n-Cl-Me	0.8110	7.1461	1.3893	-2.6480	7.7125
c-Cl-Me	0.8242	7.0019	-0.7432	-3.7318	7.6144
n-Cl-Cl	1.0097	8.2466	-0.1860	-0.1860	8.2466
c-Cl-Cl	1.0232	8.1053	-1.1891	-1.1891	8.1053

Figure 6.7: Relationship between increasing xc+yc and xc2,yc2,xy2 on the flanking ring

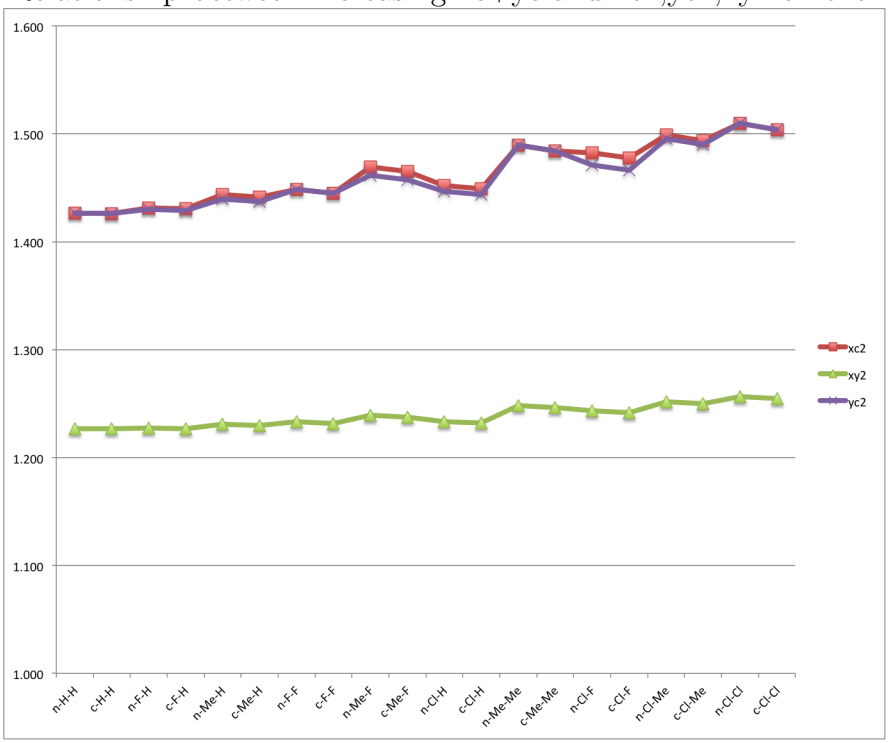


Table 6.13: Ranking X-Y distance

subst XY	xy
n-Cl-H	1.7184
c-Cl-H	1.7625
n-Me-H	1.7982
n-F-H	1.8296
c-Me-H	1.8541
c-F-H	1.9254
n-F-F	1.9359
n-Cl-F	1.9445
c-Cl-F	1.9692
c-F-F	1.9764
n-Me-F	2.0010
n-Cl-Cl	2.0082
n-H-H	2.0108
c-Cl-Cl	2.0256
c-Me-F	2.0334
n-Cl-Me	2.0455
c-Cl-Me	2.0686
c-H-H	2.1300
n-Me-Me	2.1637
c-Me-Me	2.1917

Figure 6.8: xc + yc vs. the distance between X and Y

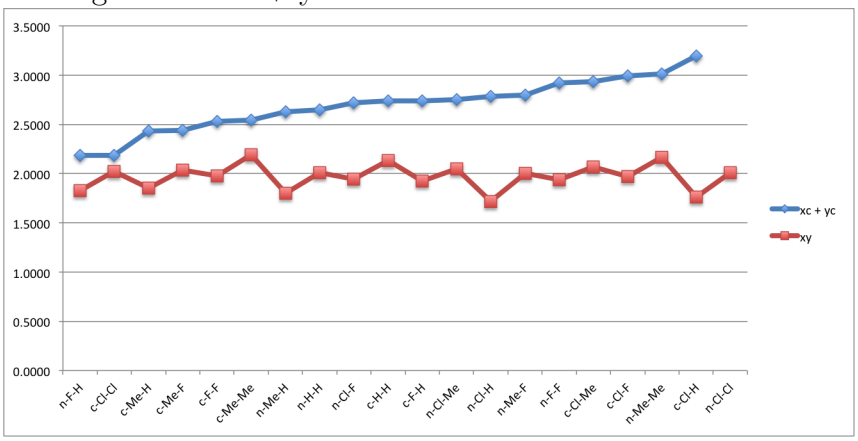
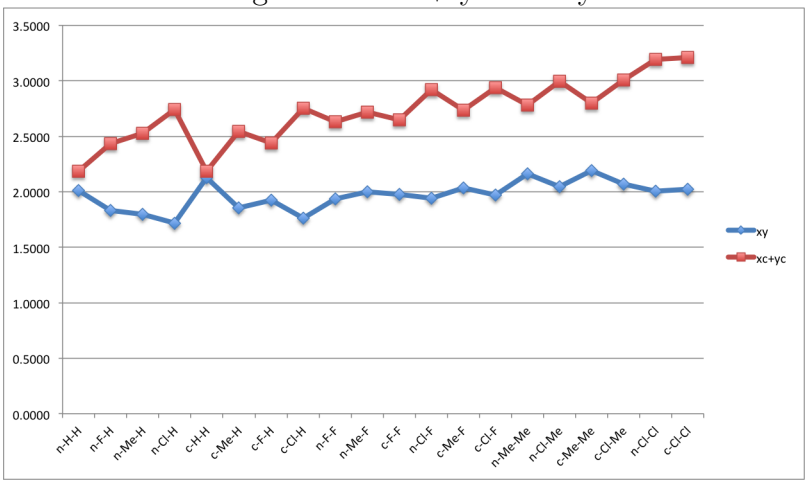


Table 6.14: xy and xc+yc ranked according to increasing total

subst. x-y	xy	xc+yc	total
n-H-H	2.0108	2.1837	9.4383
n-F-H	1.8296	2.4344	9.4562
n-Me-H	1.7982	2.5297	9.4761
n-Cl-H	1.7184	2.7393	9.5192
c-H-H	2.1300	2.1865	9.5615
c-Me-H	1.8541	2.5425	9.5666
c-F-H	1.9254	2.4409	9.5688
c-Cl-H	1.7625	2.7543	9.6086
n-F-F	1.9359	2.6307	9.6341
n-Me-F	2.0010	2.7213	9.7150
c-F-F	1.9764	2.6464	9.7169
n-Cl-F	1.9445	2.9207	9.7850
c-Me-F	2.0334	2.7362	9.7887
c-Cl-F	1.9692	2.9368	9.8567
n-Me-Me	2.1637	2.7845	9.8614
n-Cl-Me	2.0455	2.9948	9.9071
c-Me-Me	2.1917	2.7991	9.9203
c-Cl-Me	2.0686	3.0107	9.9667
n-Cl-Cl	2.0082	3.1934	9.9850
c-Cl-Cl	2.0256	3.2097	10.0472

Figure 6.9: xc + yc and xy



Correlation Analysis

Much information regarding how the variables covary can be ascertained just by looking at the correlation matrix itself. The diagonal elements of a correlation matrix are all 1 because each variable will correlate perfectly with itself. The off-diagonal elements are the correlation coefficients between pairs of variables. If there are clusters of large correlation coefficients between certain subsets of variable then that subset could be measuring the same qualities of an underlying dimension. These underlying dimensions are called factors or latent variables in PCA and/or PLS. PCA and PLS analyses will be examined in the following sections. For now, a visual analysis of the correlation matrix (Table 6.10) of the original and normalized variables reveals that all of the variables except for xy and the angle2 and angle3 (the bend of the flanking ring) are highly correlated with each other as well a(> 0.900) with xc+yc. Due to the symmetry of the molecule, it is expected that the xc2, yc2, a2, b2 and angle1 and angle4 (on the flanking ring) would be highly correlated. In the simple geometric analysis, it was seen that all these variables had a direct relationship to xc+yc. The variables angle2 and angle3 indicate the bend of the flanking Carbon axis at the point of the Carbon-Carbon triple bond and the values are not correlated with xc+yc (discussed in the simple geometric analysis).

Part and Partial Correlations

The correlation matrix reveals that many of the independent variables have high multiple correlations with each other; yet,they may be only weakly correlated with the dependent variable. It would be informative to see the relationship between each independent and dependent variable after correcting for the correlation of the independents with each other. For example, the simple geometric analysis already showed that as xc+yc increased (by 1.026 Å), xc2 and yc2 increased by 0.08 while xy2 increased by 0.03 and the variable xy2 (which corresponds to the C-C triple bond in the flanking ring) has a strong linear relationship with xc+yc, but since xc+yc also has a strong linear relationship with xc2 and yc2, it would be useful to know if the relationship between xy2 and xc+yc is direct, or if it is derived from the relationship between xy2, xc2 and yc2.

In order to analyze how xc+yc responds to each of these variables, it is useful to look at the zero-order, partial and part correlations obtained from a linear regression. The zero-order correlation is the same as the bivariate Pearson correlation coefficient. Partial correlation is a measure of the correlation between two variables in which the effects of other variables are removed from both the other independent as well as the dependent variables. In effect,the influence of everything else in the model is subtracted out. Graphical examples of partial correlations are provided in Figure 6.12 and Figure 6.13, where the linear relationship between xc+ yc and variable xy and variable angle1, respectively, are plotted, after the influence of all other variables has been removed. Part correlation removes the effects of other independend variables from an independent variable but does not remove these effects from the dependent

Table 6.15: Part and Partial Regression Coefficients (unnormalized)

variable	Zero-Order	Partial	Part
xy	0.158	-0.952	-0.203
xc2	0.933	-0.634	-0.054
yc2	0.910	-0.378	- 0.027
xy2	0.907	-0.731	-0 .070
c2	0.913	0.825	0.096
angle1	0.598	0.803	0.088
angle2	-0.273	0.790	0.085
angle3	-0.363	-0.708	-0.066
angle4	0.534	-0.708	-0.066
a	-0.482	0.760	0.077
b	-0.484	-0.636	-0.054
a2	0.917	-0.417	-0.030
b2	0.900	0.342	0.024

Table 6.16: Part and Partial Regression Coefficients (normalized)

variable	Zero-Order	Partial	Part
dimer_diff_xy	0.145	-0.995	-0.188
dimer_diff_xc2	0.933	0.681	0.018
dimer_diff_yc2	0.908	0.616	0.015
c2_diff	0.47	-0.275	-0.006
angle1_diff	0.959	0.594	0.014
angle2_diff	0.137	0.969	0.076
angle3_diff	-0.375	0.931	0.049
angle4_diff	0.949	0.821	0.028
a_diff	-0.483	0.369	0.008
b_diff	-0.493	-0.492	-0.011
a2_diff	0.475	0.397	0.008
b2_diff	0.436	-.457	-0.010

variable.

Below are the zero-order, partial and part correlations for the original and normalized data. Variable xy has a very low zero-order correlation with xc+yc but it has a much higher partial and part correlation. As a matter of fact, it has the highest partial and part correlation of the set of variables. This means that while xy is not highly correlated with the dependent variable when considered by itself, it is indeed correlated to the dependent variable when the influence of the other independent variables is accounted for. This finding is consistent with the linear relationship implied by the partial regression plot. The normalized data-set reflects the same trend with dimer_diff_xy having the highest partial and part correlation of the set.

Figure 6.10: Triptycene Dimer Correlation Matrix (Original Variables)

	a	xy	b	a2	xc2	xy2	yc2	b2	angle1	angle2	angle3	angle4
a	1.000	-.383	.986	-.982	-.988	-.981	-.982	-.971	-.947	-.282	.275	-.958
xy	-.383	1.000	-.401	.373	.446	.492	.492	.423	.291	.635	.481	.374
b	.986	-.401	1.000	-.976	-.988	-.973	-.971	-.982	-.953	-.233	.236	-.974
a2	-.982	.373	-.976	1.000	.975	.966	.968	.993	.906	.245	-.295	.924
xc2	-.988	.446	-.988	.975	1.000	.994	.990	.973	.945	.264	-.254	.963
xz2	-.981	.492	-.973	.966	.994	1.000	.997	.962	.915	.320	-.205	.935
yc2	-.982	.492	-.971	.968	.990	.997	1.000	.962	.920	.304	-.190	.933
b2	-.971	.423	-.982	.993	.973	.962	.962	1.000	.899	.251	-.233	.929
angle1	-.947	.291	-.953	.906	.945	.915	.920	.899	1.000	.073	-.311	.986
angle2	-.282	.635	-.233	.245	.264	.320	.304	.251	.073	1.000	.289	.178
angle3	.275	.481	.236	-.295	-.254	-.205	-.190	-.233	-.311	.289	1.000	-.292
angle4	-.958	.374	-.974	.924	.963	.935	.933	.929	.986	.178	-.292	1.000

Figure 6.11: Triptycene Dimer Correlation Matrix (NormalizedVariables)

	a_diff	xy_diff	xc_plus_yc_diff	b_diff	a2_diff	xc2_diff	xy2_diff	yc2_diff	b2_diff	angle1_diff	angle2_diff	angle3_diff	angle4_diff
a_diff	1.000	-.664	-.954	.982	-.996	-.987	-.988	-.985	-.980	-.959	-.580	.042	-.956
xy_diff	-.664	1.000	.426	-.668	.711	.688	.696	.716	.734	.618	.619	.436	.646
xc_plus_yc_diff	-.954	.426	1.000	-.948	.937	.936	.930	.917	.922	.929	.458	-.172	.924
b_diff	.982	-.668	-.948	1.000	-.988	-.985	-.974	-.968	-.994	-.977	-.503	.003	-.984
a2_diff	-.996	.711	.937	-.988	1.000	.991	.989	.987	.992	.960	.592	.002	.964
xc2_diff	-.987	.688	.936	-.985	.991	1.000	.995	.989	.984	.978	.509	-.049	.974
xy2_diff	-.988	.696	.930	-.974	.989	.995	1.000	.997	.976	.963	.531	-.019	.952
yc2_diff	-.985	.716	.917	-.968	.987	.989	.997	1.000	.972	.958	.533	.014	.943
b2_diff	-.980	.734	.922	-.994	.992	.984	.976	.972	1.000	.961	.564	.058	.975
angle1_diff	-.959	.618	.929	-.977	.960	.978	.963	.958	.961	1.000	.361	-.122	.986
angle2_diff	-.580	.619	.458	-.503	.592	.509	.531	.533	.564	.361	1.000	.200	.436
angle3_diff	.042	.436	-.172	.003	.002	-.049	-.019	.014	.058	-.122	.200	1.000	-.123
angle4_diff	-.956	.646	.924	-.984	.964	.974	.952	.943	.975	.986	.436	-.123	1.000

Principle Component Analysis

In addition to examination of the correlation matrix and the part and partial correlations, a Principle Component Analysis (PCA, Introduced in Chapter 3) was done to more formally identify groupings for the independent variables. Only 2 components were required to explain 92% of the variance for both the original and normalized data sets. Tables 6.17 and 6.18 shows that all the variables except for xy, angle2 and angle3 load most heavily on the first component which explains about 75 percent of the variation in the data, while angle2, angle3 and xy load

Table 6.17: Principle Components (Original Data)

variable	1	2
a	-.992	.049
xy	.459	.803
b	-.992	.047
a2	.980	-.071
xc2	.997	-.021
xy2	.990	.047
yc2	.989	.047
b2	.979	-.022
angle1	.946	-.189
angle2	.291	.751
angle3	-.242	.801
angle4	.968	-.105

Table 6.18: Principle Components (Normalized Data)

variable	1	2
a_diff	-.990	.067
xy_diff	.744	.534
b_diff	-.987	.070
a2_diff	.998	-.019
xc2_diff	.992	-.091
xy2_diff	.989	-.057
yc2_diff	.988	-.025
b2_diff	.994	.017
angle1_diff	.963	-.211
angle2_diff	.581	.461
angle3_diff	.019	.908
angle4_diff	.969	-.176

heavily on component 2, which explains an additional 15 percent of the variation.

The groupings obtained by PCA are consistent with the groupings implied by examination of the correlation matrix and the trends seen from the simple geometric analysis.

Figure 6.12: Partial Regression Plot: xc_plus_yc vs dimer_diff_xy

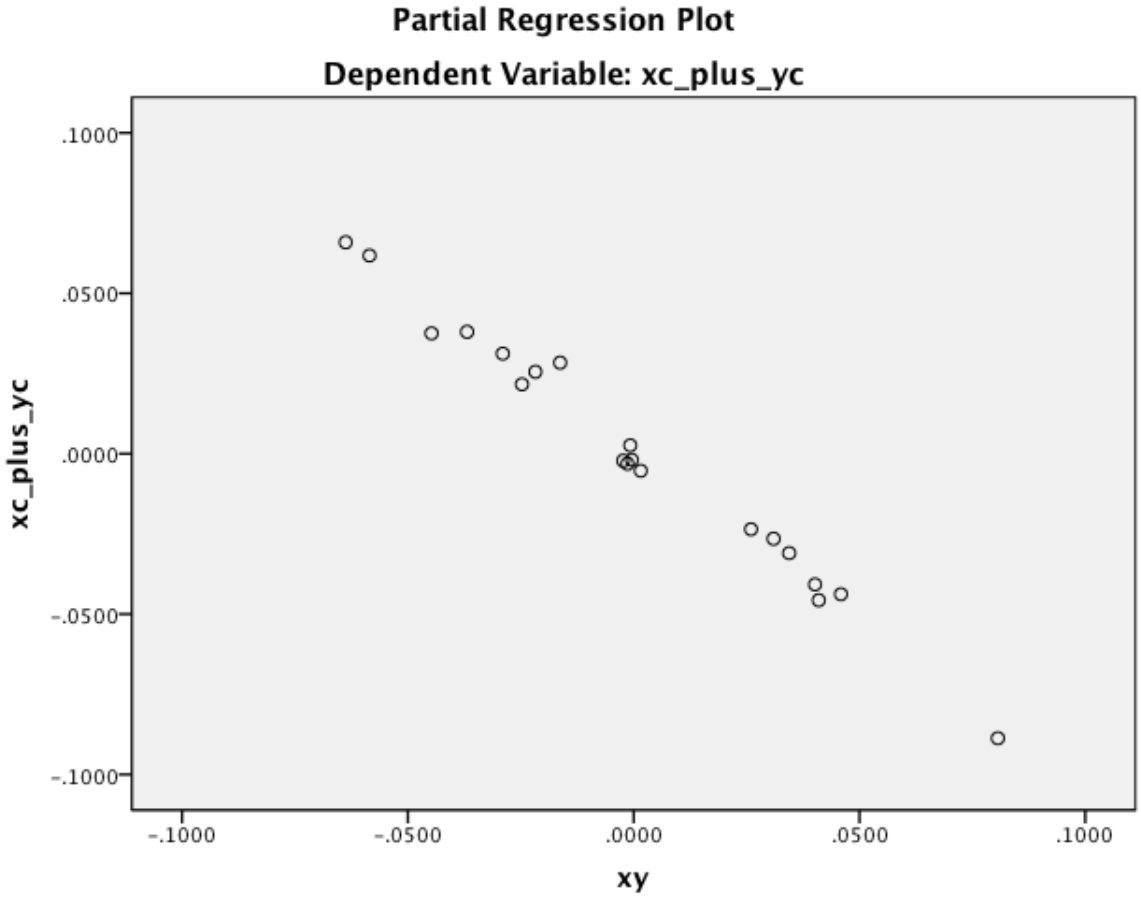
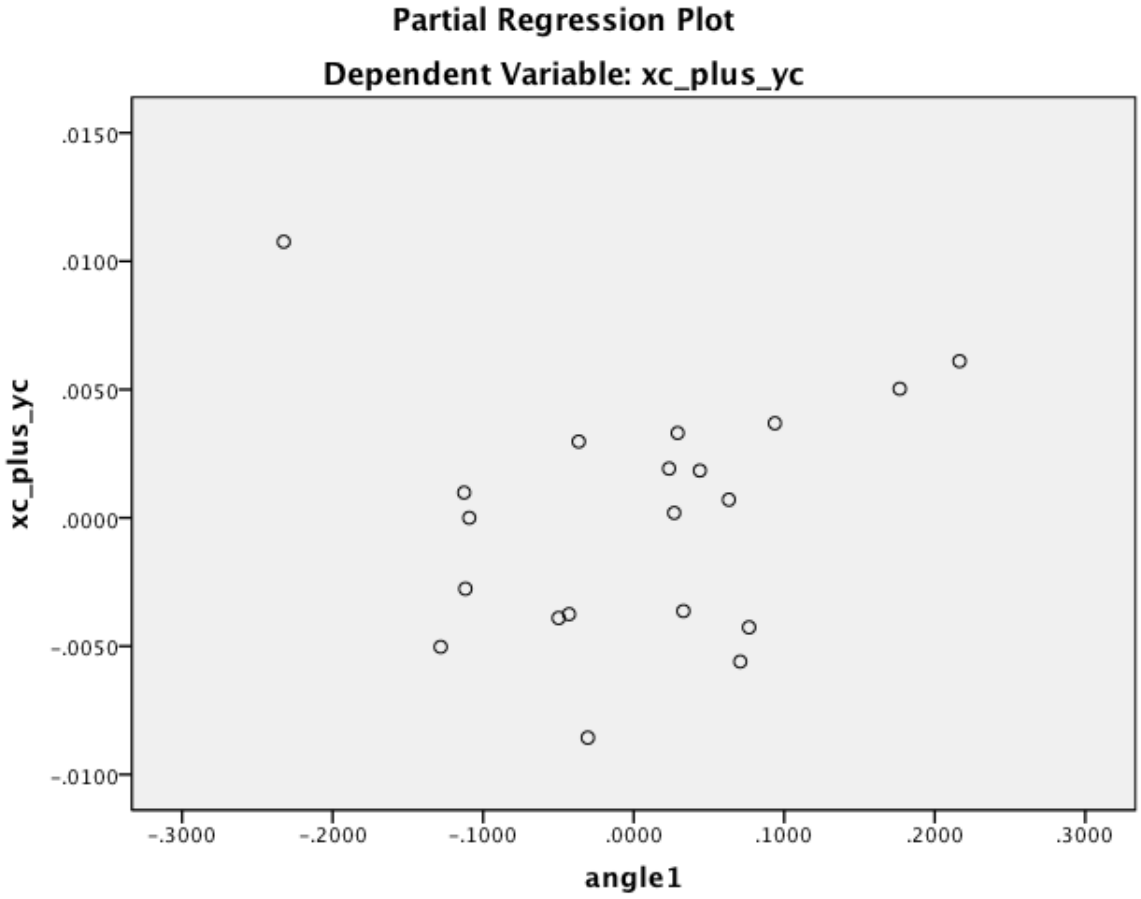


Figure 6.13: Partial Regression Plot: xc_plus_yc vs c2_diff



Linear Regression Analysis

Regression analysis in a model having many more independent variables than observations, as is the case with the triptycine data, generally results in an overfit and unreliable model. Therefore, a stepwise regression model was used to select an subset of best predictors from amongst the independent variables. Stepwise regression starts by selecting the variable having the greatest correlation with xc+yc (which happens to be a or a_diff) and then it selects the variable to add that has the largest partial correlation with xc+yc after variance from a is accounted for (xy in this case). When either a stepwise or forward regression method is used on the original and normalized data sets, very good results (R^2=0.910) can be obtained using only one variable, a or a_diff respectively. The addition of variable xy (or xy_diff) slightly improves the predictive power of the model (R^2=0.986). This confirms the results from the partial regression plot that showed xy to have strong relation with xc+yc. One might ask why is variable xy and not another variable having higher simple correlation with xc+yc selected by the stepwise regression method? The qualitative answer is that most of these variables are also highly correlated with variable a, and carry much of the same information. So, after contribution of variable a is removed, there is not much these highly correlated variables have left to explain. This means to that variable a does a very good job by itself in explaining about 91 percent of the variation in xc+yc, when xy is added to the model,then xy does a reasonably good job in explaining an additional 5 percent. Many of the variables that are highly correlated with variable a could also be used to predict xc+yc, angle1 for example, could be used just as well and may predict as well as variable a.

Partial Least Squares

Partial Least Squares (PLS) was also used to explore the relationships between the data. PLS aims to derive factors(components) of closely related(highly correlated) variables. As was mentioned earlier, PCA can be used to identify underlying dimensions (factors) that can be used as a way to reduce the data set from a group of interrelated variables into a smaller set of uncorrelated factors. This allows for a maximum amount of common variance to be explained with the smallest number of factors. Often, it is possible to interpret these extracted factors in terms of the underlying physical system. Similarly PLS can be used as a variable reduction technique. and be interpreted in a similar fashion. PLS differs from PCA because *both* the independent variables (X) and the dependent variables or (Y) are reduced to principal components. The PCA algorithm is only concerned with maximizing variance in the independent variables and does not consider the response at all.

To construct the principle components of the independent variables(X) the PLS algorithm iteratively maximizes the strength of the relation between successive pairs of the X and Y component scores by maximizing the covariance of each X-score with the Y variables. The methodology employed in the PLS algorithm gives PLS some advantages over Multiple Linear Regression type methods: 1) The X components used to predict Y will be orthogonal so the method will not fail if the original X variables are multi-collinear, 2) For the Regression only a few of the components are used in the prediction so PLS will not overfit the data even when there are more variables than cases/observations.

In the following PLS analysis, the full set of original variables is explored (xy, xc2,yc2,xy2,angle1,angle2,angle3,angle4,a,b,a2 and b2) to determine how xc+yc varies with respect to the other variables. Since here PLS is used as an exploratory technique rather than strictly as a tool for prediction, it is instructive to look at xc+yc as both the dependent and independent variable. The objective is to explore how the system responds to changes in xc+yc (where xc+yc will be the independent variable) as well as how changes in the system are reflected in xc+yc (where xc+yc will be the dependent variable).

PLS with xc+yc as the dependent variable

First, xc+yc is used as the dependent variable and the remaining variables are used as independent variables. The Table 6.19 below has the latent factors listed by row. The cumulative Y variance is the percent of variance in the Y variable accounted for by the latent factors and likewise the cumulative X variance explains the variance in the X factors. The 1st component explains the bulk of the variation in Y (0.894)as well as X (0.751). With the 2nd (Y:0.048, X:0.157) and 3rd component (Y:0.054, X:0.022) contributing far less. The first three factors explain virtually all the variance in Y (0.996). The more a factor explains the variation in Y, the more likely it will be to explain the variation in a new sample of dependent variables and the more a factor explains the variation in the X variables the better it reflects the observed

Table 6.19: Proportion of Variance Explained (PLS with xc+yc as the dependent variable)

Proportion of Variance Explained

Latent Factors	Statistics				
	X Variance	Cumulative X Variance	Y Variance	Cumulative Y Variance (R-square)	Adjusted R-square
1	.751	.751	.894	.894	.889
2	.157	.908	.048	.942	.935
3	.022	.930	.054	.996	.995
4	.020	.950	.003	.999	.998
5	.043	.992	.000	.999	.999

values of the set of independent variables. Given that Factor 1 explains most of the variation in X and Y, it alone could be used both for prediction (since factor 1 explains the variability in Y) and for explaining the relationships between the independent variables (since factor 1 explains the bulk of the variability in X). No conclusions can be made about *which* variables are significantly effected by changes in xc+yc yet, but the fact that 1 latent factor explains the bulk of the variance indicates that the response (xc+yc) could be explained by fewer variables. Unlike the stepwise linear regression technique which filters out variables, PLS uses all the independent variables, including those having low correlation with the dependent variable. PLS is known not to perform as well as regression as an explanatory technique because it does not filter out variables of minor causal importance. [18]

The coefficients given in the"Variable Importance in the Projection" reflect the relative importance of each X variable for each X factor in the prediction model. Since the Y-scores are predicted from the X-scores, the VIP coefficients represent the importance of each X variable in fitting both the X and Y scores. It is customary to remove independent variables whose VIP coefficient is < 0.8 and also have a small regression coefficient from the model. Based on the VIP coefficients in the table below, xy, angle2, angle3 and possibly also a and b could be removed. Before deciding definitively which variables could be removed, the regression coefficient matrix should be examined. This does not mean that xy, angle2 and angle3 are not explanatory for xc+yc, but the PLS just shows that xy doesn't group well in a factor or component.

The tables of loadings and weights indicate how much each independent variable contributes to the latent factor in each column. X-weights and X-loadings are similar and serve similar interpretive uses. The Weights in Table 6.21 are the X-weights, representing the correlation of the X variables with the Y-scores. The Loadings table (Table 6.22) shows the X-loadings and can be thought of as the directions of the lines for each independent variable in X-space. As with PCA, the loadings in PLS are sometimes used to infer meaning to the factors but this may be confusing if there are cross-loadings on more than one factor. There are no rough trends in the Weights or Loadings table that can be used as a basis for categorizing or imparting meanings for the factors as was done with the PCA analysis. Sometimes a cut-off is used (0.25-0.70 depending on the goal of the analysis) to determine whether an independent variable will be included in

Table 6.20: Variable Importance in the Projection

Variable Importance in the Projection

Variables	Latent Factors				
	1	2	3	4	5
a	1.175	1.147	1.124	1.123	1.122
xy	.194	.682	.881	.881	.880
b	1.164	1.135	1.107	1.106	1.106
a2	1.125	1.096	1.072	1.074	1.074
xc2	1.143	1.114	1.084	1.082	1.082
xy2	1.115	1.087	1.057	1.056	1.056
yc2	1.113	1.085	1.056	1.055	1.055
b2	1.104	1.077	1.057	1.060	1.061
angle1	1.174	1.158	1.148	1.150	1.150
angle2	.185	.306	.472	.472	.473
angle3	.435	.496	.625	.624	.625
angle4	1.163	1.138	1.118	1.119	1.119

Cumulative Variable Importance

Table 6.21: Weights

Weights

Variables	Latent Factors				
	1	2	3	4	5
a	-.339	-.097	-.172	.024	-.033
xy	.056	-.839	-.721	.171	.033
b	-.336	-.068	-.106	.051	.146
a2	.325	-.011	-.140	-.517	-.449
xc2	.330	-.004	-.024	.073	.128
xy2	.322	-.053	-.032	.085	.270
yc2	.321	-.055	-.040	.039	.143
b2	.319	-.061	-.181	-.543	-.564
angle1	.339	.229	.279	.496	.107
angle2	.054	-.316	.455	-.100	.572
angle3	-.125	-.332	.493	.022	-.609
angle4	.336	.138	.193	.417	.143
xc_plus_yc	.317	.173	.458	.123	.029

a factor. Even when using 0.25 as cut-off factor (acceptable since PLS is being used as an exploratory technique here) there still is not any clear membership to any category with many of the variables being cross-loaded between Factor 1 and Factor 2.

Table 6.23, "Parameters", gives the Regression parameter (coefficient) estimates. These are the regression coefficients and indicate the rate of change of the dependent variable per unit change in the independent variable (listed on the left). From this table, variables that don't have much influence on changes in xc+yc can be identified. It can be seen that angle1, angle2, angle3 and angle4 have little influence on xc+yc. The PLS results could be refined by eliminating variables with small regression coefficients *if* they also have a smaller coefficient for the "Variable Importance in the Projection" (discussed previously). The variables that meet both criteria are angle2, angle3. Variable xy had a low VIP coefficient but a fairly high Regression coefficient parameter (-0.967) . Another thing to keep in mind when evaluating variables for further study (or elimination) is that he physical aspects of the model should be maintained. Especially since the goal here is not to develop a predictive model but rather to explore the variable space. Given

Table 6.22: Loadings

Loadings

Variables	Latent Factors				
	1	2	3	4	5
a	-.333	-.008	-.141	.084	-.026
xy	.139	-.690	-.420	.345	-.008
b	-.333	-.002	-.090	.001	.102
a2	.329	.014	-.045	-.410	-.076
xc2	.334	-.027	-.036	.027	-.004
xy2	.330	-.080	-.026	-.042	.020
yc2	.330	-.080	-.025	-.026	-.024
b2	.328	-.027	-.062	-.367	-.159
angle1	.320	.118	.101	.563	-.077
angle2	.085	-.572	.623	-.567	.746
angle3	-.094	-.589	.647	.309	-.724
angle4	.325	.048	.065	.449	-.001
xc_plus_yc	1.000	1.000	1.000	1.000	1.000

Table 6.23: Parameters

Parameters

Independent Variables	Depend...
	xc_plus_yc
(Constant)	-13.493
a	-.808
xy	-.967
b	-.607
a2	-.323
xc2	1.094
xy2	2.740
yc2	.866
b2	-.583
angle1	.039
angle2	.031
angle3	.024
angle4	.028

this, it would not be advisable to remove all of the angle variables since the interest is in how xc+yc responds to changes in the angle even if the regression coefficient and the VIP coefficients indicated otherwise.

PLS with xc+yc as the independent variable

The PLS procedure was repeated with xc+yc as the independent variable in order to give the rate of change of each variable holding all other variables constant. The variable xc+yc alone can account for 66 % of the variance (Table 6.24). To help understand the relationship of the variables to xc+yc, or equivalently, to see how the variables change with xc+yc, beta coefficients can be used. These values are provided along with a regression constant in the table of parameters in the PLS output. (Table 6.24) gives each variable (a, b, xy, angle1, etc) as a linear function of xc+yc. For each variable there is a constant and a regression (beta) coefficient. Angle 4, for example, would be estimated by angle4=100.088 + 9.398*xc+yc. The constant (100.088 for angle4) is the value of thedependent variable (angle4 say) when xc+yc is equal to zero. This means that, for angle4, when holding all other variables constant, a change in xc+yc by one unit would result in a 9.398 change in angle4. And so on for other variables using their respective beta coefficients. This is same result that would be obtained if each of these variables is regressed as dependent variables on xc+yc one at a time (the constant and beta coefficient would be the same as PLS). The weights for the PLS when xc+yc is the independent variable are the correlations of these variables with the one factor (which is a factor with one variable, xc+yc). Also note that this is just same as simple correlation of the variable with xc+yc.

Table 6.24: Proportion of Variance Explained (xy)

Proportion of Variance Explained

Latent Factors	Statistics				
	X Variance	Cumulative X Variance	Y Variance	Cumulative Y Variance (R-square)	Adjusted R-square
1	1.000	1.000	.665	.665	.646

Table 6.25: Parameters for Proportion of Variance Explained (xy)

Parameters

Independent Variables	Dependent Variables											
	angle4	angle3	angle2	angle1	b2	yc2	xy2	xc2	a2	b	xy	a
(Constant)	100.088	182.329	175.453	102.747	2.586	1.224	1.152	1.218	2.597	3.180	1.775	3.162
xc_plus_yc	9.348	-1.708	.854	8.407	.117	.087	.032	.090	.113	-.245	.071	-.238

Multilayer Perceptron Neural Net

A Multilayer Perceptron (MLP) Neural Net Architecture was also used to investigate the influence of xc+yc on the overall strain of the system. The output layer consisted of the single variable xc+yc and the input layer consisted of the remaining variables as shown in Figure 6.14. The number of neurons in the hidden layer was determined automatically to optimize network performance. The network was trained until the relative error was below a specified cut-off. The same procedure was repeated for the normalized data set. Figures 6.15 and 6.16 show the

results for the normalized importance. The relative importance is calculated by the degree to which the error increases when a particular variable is removed from the MLP neural network. The MLP results were similar to the linear regression results with xy and angle1 being the most important in the original data set and xy_diff being the most important and angle1_diff angle4_diff (equivalently ranked) were the most important. It is important to note that the results of the sensitivity analysis are slightly different each time the neural net is executed. This is due to the fact that the procedure uses random number generation during random assignment of partitions, random subsampling for initialization of synaptic weights, random subsampling for automatic architecture selection, and the simulated annealing algorithm used in weight initialization and automatic architecture selection. However the results were not significantly different from run to run, with only the ordering of the variables being slightly affected.

Figure 6.14: Multilayer Perceptron NN Architecture

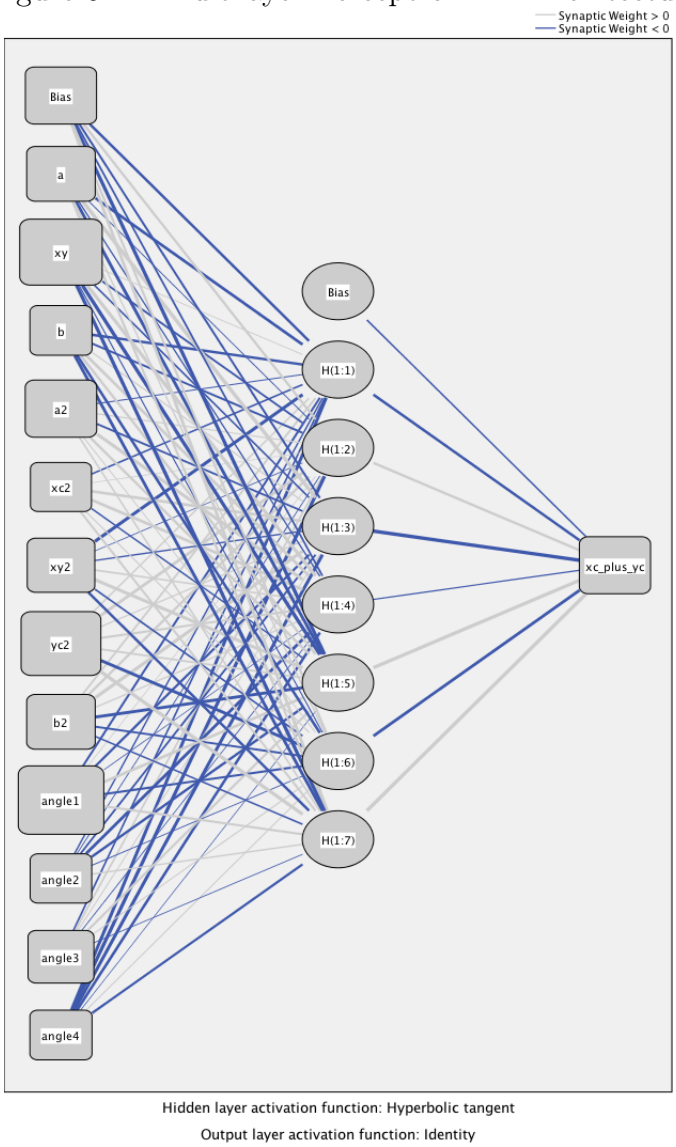


Figure 6.15: MLP Variable Importance for Original Data Set

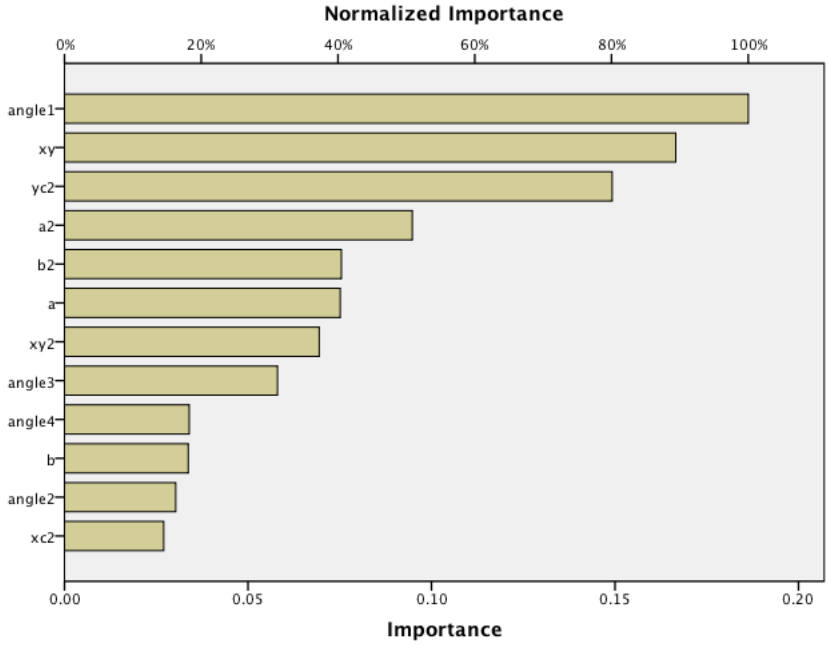
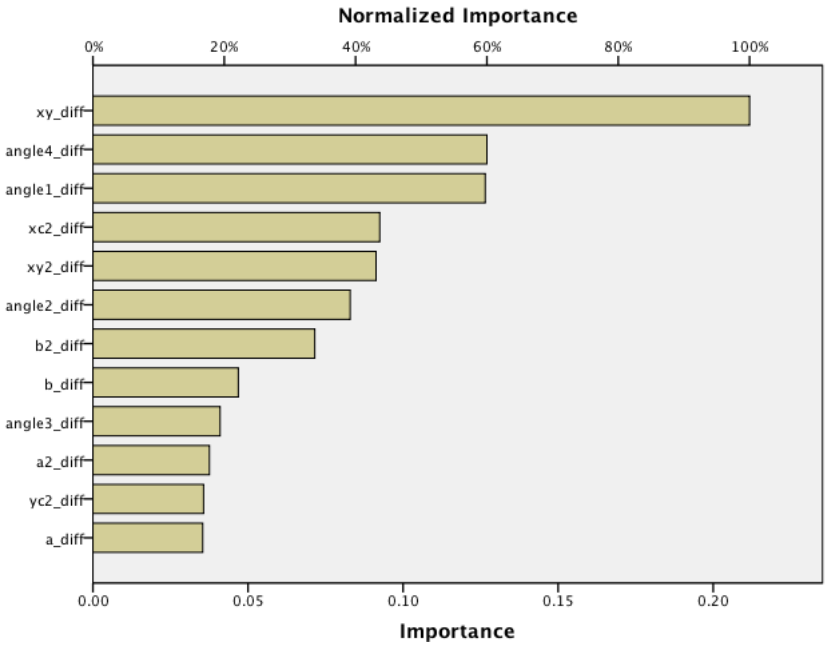


Figure 6.16: MLP Normalized Variable Importance for Normalized Data Set



6.4 Discussion and Conclusions

If the goal of this analysis were to form a simple predictive model, linear regression could be used with a single variable. The stepwise regression selected variable a (or angle1 or angle4) as the best variable but since almost all the variables are highly correlated any of the other variables could be used (with the exception of angle2, angle3 and xy which did not have a linear relationship with any of the other variables) to model the changes in xc+yc with a good value of R^2. The stepwise regression also selected xy (and xy_diff) as a second variable in the linear regression, but xy only had a small contribution to the overall predictive power of the model It is noteworthy, however, that these dependent variables are have strong linear, and possibly more complex nonlinear associations with each other which may not be adequately captured in a linear regression model.

The MLP neural network sensitivity analysis confirmed that xy and angle1 (or angle4) were the most important variables for the original and normalized data sest. Discrepancies between the results from a neural network and linear regression could be partially attributed to the fact that the datasets are heteroscedastic (increasing or non constant variance). One of the assumptions of linear regression is that the errors would have the same distribution (normal distribution with zero mean and common variance) for each observation. This can be seen by examining the residual plot of xc+yc versus the predicted value and noting that the spread increases as xc+yc gets larger. Neural networks do not make this assumption. However since the results were similar between the NN and the linear regression, it can be assumed that the regression model has not been seriously degraded by the heterocedasticity.

For this study, the main concern is how the system changes as strain (as indicated by an increase in the lengths of xc+yc) is varied. So, in this case, xc+yc is modeled to be the independent variable and all the others to be dependent variables. This relationship was examined using PLS; angle1 and angle4 were found to be most strongly influenced by changes in xc+yc. Overall, this analysis suggests that in the actual physical system all of these variables are strongly connected in a complex way that is not adequately described in traditional linear models such as MLR and PLS. An alternative approach using structural equation modeling or a dynamic system model would be productive for follow on study.

Chapter 7

Application of QM Descriptors to Fingerprinting 5HT2a Fingerprints

7.1 Introduction

Speculation based on Similarities of Serotonin to Psychoactive Drugs

Much of what is currently known about the chemical and biochemical mechanisms that underly affective disorders stem from early speculation on the similarities between hallucinogenic drugs and the serotonin molecule itself.[19] In 1943, the hallucinogenic properties of the synthetic ergoline compound LSD (d-lysergic acid diethylamide) was serendipitously discovered by chemist Albert Hoffman. Five years later, in 1948, serotonin was purified from approximately '900 liters of serum collected from almost two tons of beef blood'.[20] Then, in 1953, during a routine survey of various tissues, relatively high concentrations of 5-HT were found in brain.[21] Shortly thereafter, based on the observation that LSD could antagonize 5-HT in peripheral tissues, plus the structural similarity between these two indole-containing structures, Woolley and Shaw[22] first proposed that the 'mental disturbances caused by lysergic acid diethylamide were to be attributed to an interference with the action of serotonin in the brain' in their 1954 paper 'A Biochemical and Pharmacological Suggestion about Certain Mental Disorders'.[23] The serotonin hypothesis was later extended to include simple indoleamine hallucinogens as psilocin, which are close structural analogs of 5-HT (Figure 7.2) and the phenethylamine hallucinogens, such as mescaline. Interest in the role of serotonin in behavior began to rapidly increase and this acceleration of interest has continued to the present time.

5-HT is classified into seven subtypes based on their structural and functional characteristics. Most of what is known regarding the subtypes occurred mainly through radio-labeled ligand binding before receptor cloning technology was available.[24] Many clinical studies have confirmed that the 5HT2a subtype plays a critical role in the action of both hallucinogens and antipsychotic medications.[25] While it is now known that there are certainly many other receptor systems involved, the serotonin hypothesis of schizophrenia has been of considerable value in leading to the development of many of the early antipsychotic medications. The ability of a potential drug

Figure 7.1: Serotonin and LSD

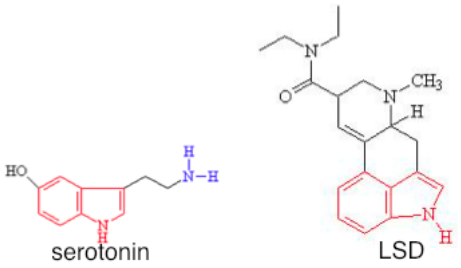
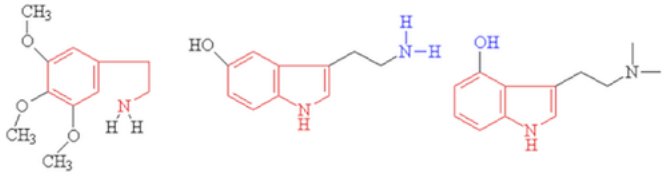


Figure 7.2: Mescaline, Serotonin and Psilocin

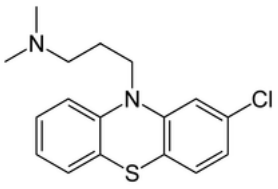


to block LSD activity was used to identify its potential as an antipsychotic. Thus, understanding the mechanism of hallucinogens continues to have the potential to provide important clues about the basis for psychosis in this disease. The basic pharmacophore for the interaction of hallucinogens with the serotonin receptor was in place long before any information about the structure of the receptor was known.[26]

In 1950 Thorazine (chlorpromazine) was synthesized and was the first drug developed with specific antipsychotic action. The introduction of chlorpromazine into clinical use has been described as the single greatest advance in psychiatric care.[27] The availability of antipsychotic drugs greatly diminished the use of electroshock treatments and psychosurgery, and helped drive the deinstitutionalization movement. Many other drugs were developed with the same tricyclic phenothiazine scaffold of Thorazine, but the mechanism of action was still unknown. Thorazine (and most other typical antipsychotics) were later found to cause serious side effects such as tardive dyskinesia which encouraged research into the mechanism of therapeutic antipsychotic action while curtailing unwanted side effects. A Science paper in 1959 was published with a two line abstract that simply stated, "Chlorpromazine is shown to be a powerful electron donor. Observations are described supporting the assumption that the therapeutic action of this drug is connected with this property." This period of time marks the beginning of a barrage of research attempting to determine the mechanism of therapeutic antipsychotic drug action efforts to reduce side effects and maximize therapeutic value.

It is quite an oversimplification to attribute the therapeutic activity of a particular drug to a single receptor. The general class of receptors (G-Protein Coupled Receptors or GPCRs) which are the target of most CNS drugs consist of a family of very structurally similar proteins and subtypes. For example, the commonly prescribed atypical antipsychotic drug Olanzapine (

Figure 7.3: Thorazine



Zyprexa ^{TM}) has the histamine H1 receptor as the primary target (K_i =3.5 nM) but also has high affinity for the 5HT2a receptor (K_i = 1.9 nM) and 5HT2c receptor (K_i = 7.1 nM) as well as a number of other GPCR targets: the Dopamine D1 receptor (K_i = 250 nM), the Dopamine D3 receptor (K_i = 54 nM), the Dopamine D_4.2 receptor (K_i = 28 nM), α-2 adrenoceptor (K_i = 230 nM) and finally the Muscarinergic receptor (K_i = 26 nM). The "activity" that is observed is typically the result of a drug interacting with many different receptors and receptor subtypes.

One approach to simplifying the problem, is to restrict the analysis to one subtype at a time and investigate the mechanistic importance of that subtype individually. Experimentally, this means finding a compound selective for one receptor subtype, since cross-receptor interactions make it nearly impossible to ascribe a particular effect to a particular receptor. AMDA (9-(aminomethyl)-9,10-dihydroanthracene depicted in Figure 7.4) was synthesized as a highly selective compound for 5HT2a. While it does not have a known therapeutic use, its study has led to insights regarding the involvement of the 5HT2a receptor on the mechanism of similar drug classes that do not have this selectivity: namely, the tricyclic antidepressants (Figure 7.5) and phenothiazine (Figure 7.7) antipsychotics. Both classes of agents are tricyclic amines consisting of two aromatic groups flanking a nonaromatic central ring that bears an alkylamino substituent as in AMDA.

AMDA is also thought to share a binding mode with the 2,5-Dimethoxy-4-bromoamphetamine (DOB) like antagonists. DOB is a well studied scaffold for phenylethylamine hallucinogenic activity and a 5HT2a agonist. Early on, the pharmacophore for affinity in the phenylethylamine hallucinogens was determined to be a 2,5 methoxy substitution pattern on the basic phenylethylamine ring, with presence and nature of a 4 position substituent modulating the affinity.[28] More recently, it was discovered that if the 4 position substituent was substantially bulky, the 2-5 methoxy pattern was no longer required for affinity and that in fact the molecule behaved as an antagonist once the 4 position substituent becamse substanstially bulky.[29] This finding could potentially lead to a better understanding of what factors contribute to agonist vs. antagonist functionality and work has already been done in this direction. For example, in 2008, Parker et al.[30] correlated lipophilicity with the functional activity for 5HT2a receptor ligands with future work planned to design an experimental methodology for distinguishing agonists from antagonists in cases where functional activity of ligands cannot easily be measured directly. Examples such as this demonstrate the potentially wide applicability of

understanding the mechanism behind the selectivity and functional activity of selective ligands such as AMDA. With the advent of site-directed mutagenisis and cloning technology it has become easier to determine the affinity of a particular compound for a receptor but much more difficult to determine the functional activity. This is particularly true for affective disorders where it activity is difficult to assess in animal models and not always feasible to test in human models until safety screening has been performed. It was by the virtues of its relatively simple scaffold, 5HT2a selectivity and extensibility to many known drug classes that AMDA chosen as the test case for the methodologies developed in this thesis.

Figure 7.4: AMDA

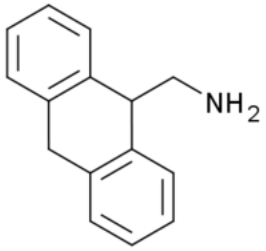


Figure 7.5: Imipramine

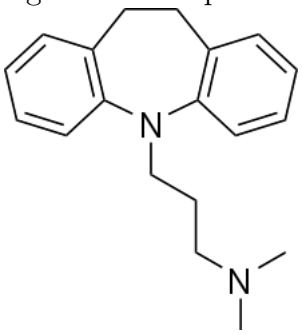
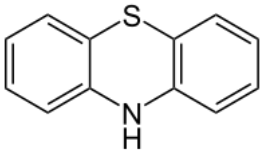


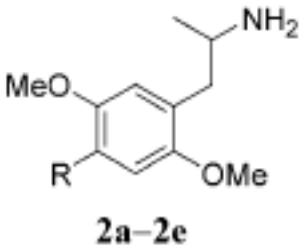
Figure 7.6: Phenothiazine



5HT2a Agonist and Antagonist Pharmacophore

What is currently known regarding the possible binding modes of ligands and the 5HT2a receptor itself was derived from a variety of methods. Molecular modeling studies are typically approached from either consideration of only the ligands (ligand-ligand approaches) or modeling interactions between a ligand and receptor macromolecule (ligand-receptor approaches). Ligand-ligand approaches involve the comparison of properties and structural features between a series

Figure 7.7: "DOB-like" scaffold



of molecules. The similarities are used to infer what the receptor surface looks like, as such, these ligand-ligand approaches are sometimes referred to as "receptor mapping". Many methods which aim to perform a three-dimensional comparison for a series of ligands require an alignment of the molecules onto a common reference frame. This can be problematic when the binding configuration is unknown.[31] . There are also comparative methods which are alignment-free and typically use a relative coordinate frame as a reference. Since the experimentally determined pharmacophore for 5HT2a antagonists is defined using relative internal distances, alignment free techniques were used for the work described in this thesis.

The basic pharmacophore (Figure 7.8) for activity for both agonists and antagonists was determined long before there was any knowledge of the receptor structure. Much of the ground work for what is currently known regarding structure and activity for CNS drugs was done using traditional statistical methods and classic methods of pharmacology. Early efforts to identify the pharmacophore for 5HT agonists (the hallucinogens) were facilitated by the use of the relatively rigid LSD molecule as a template. Similarly, for the tricyclic antipsychotic antagonists, the simple rigid tricyclic scaffold provided a convenient template. The requirements for various drug classes vary in detail, but a basic nitrogen atom and one or two aromatic groups are necessary for the majority of CNS active drugs.[32] Varieties in specific arrangements of these groups have been proposed as requirements for differences in functionality: analgesic, anticonvulsent, antidepressant, antipsychotic, hallucinogen, stimulant, etc. For example, the pharmacophore for the tricyclic antagonists Figure 7.8 has the basic arrangement of a CNS active drug but there are additional structural criteria 7.9 that are thought to differentiate the tricyclics into antipsychotics, antidepressants or sedatives.

Figure 7.8: Antagonists Pharmacophore, distances

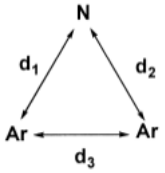
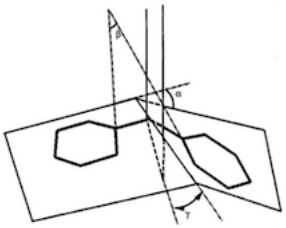


Figure 7.9: Antagonist Pharmacophore, angles



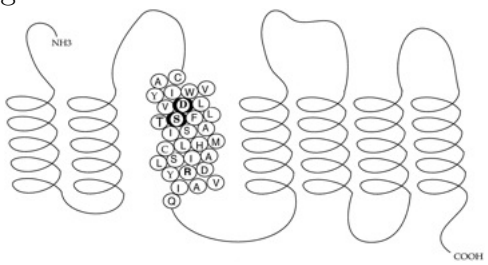
Site-Directed Mutagenisis, Homology Modeling and Docking Generated Hypotheses About Significant Interactions.

GPCRS are transmembrane proteins with a common core structure of 7 transmembrane helices. Membrane proteins are notoriously difficult to crystalize which accounts for the paucity of experimentally determined structures. A computational technique called "Homology modeling" uses a known GPCR structure as a template for an unknown structure. As mentioned, a major obstacle in understanding the interactions between 5HT2a and the ligands that interact with it that there is not a Xray crystal structure available. The human 5HT2a receptor was first cloned by Branchek et al. in 1990. In 2000 that a high-resolution crystal structure for a GPCR, bovine rhodopsin, was elucidated. There are a number of techniques that aid in using these types of data to infer the 3D structure of 5HT2a itself A technique called Homology Modeling provides a way to infer what the structure of related receptors look like using a known structure as a template based on the known genetic sequence of the receptor. Site-directed mutagenesis, a molecular biology technique in which a mutation is created at a defined site in a DNA molecule, thus causing a mutation in the expressed protein, has been used to determine specific amino acid interactions of importance. Docking is a computational method which tries to predict the preferred orientation of a ligand in a receptor binding pocket. Often, researchers use docking to explore hypotheses of a ligands binding mode using the data from site-directed mutagenisis homology models and radio-ligand binding studies. These techniques have enabled researchers together the pieces of the 5Ht2a structure puzzle that have been developing since 1943.

While there is still quite a bit of uncertainty the exact 3D structure of the various receptors, there is consensus on the role of some specific interactions, especially ones that are conserved across many members of the GPCR class. Most importantly, for 5HT2a ligands a highly conserved aspartic acid (Asp-155, indicated with the abbreviation D in Figure 7.10 and is also indicated as D155 in Figure 7.11) serves as an anchor for the terminal amine group.[33] For both agonists and antagonists, this aspartic acid serves as an anchoring point for the basic amine that seems to be a requirement for CNS activity for both agonists and antagonists. Agonists are thought to acquire a larger portion of their binding affinity through interaction with relatively hydrophilic amino acid side chains in the receptor pocket than antagonist. The antagonist binding site are thought to bind mainly through hydrophobic interactions with the receptor. By definition, "antagonists" do not exert conformation changing forces on the amino acids inside the binding pocket but rather blocks or dampens an agonist mediated response.

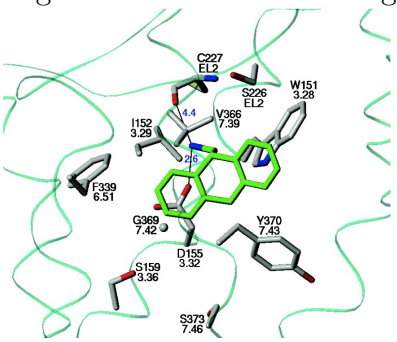
Site-directed mutagenisis and docking studies of antagonists (Figure 7.11 shows AMDA docked into the binding site) indeed confirm that hydrophobic interactions with residues in TMH2 and TMH7 stabilize the complex.[34] Antagonists with heteroatoms could form hydrogen bond interactions with three highly conserved residues in TMH3 (Cys148, Asp155, and Ser159). The relative featurelessness of AMDA and related compounds made reliable docking a problem (specifically described in the next section "Cyproheptadiene and AMDA binding mode") .

Figure 7.10: 5HT2A Conserved Residues



While the energetic interactions between individual amino acids in the receptor and the ligands are significant. In this work, the focus is solely on the 14 atom aromatic frame of the AMDA scaffold. This simplifies the problem greatly and still will allow one to gauge the interactions that stabilize the complex in a relative sense. Given the current challenges in docking approaches that utilize the structural information of the receptor, using approaches that do not rely on the receptor geometry at all are advantageous since they are not biased by experimental errors or bias.

Figure 7.11: AMDA docking



7.2 Dataset and Computational Details

AMDA is a selective high-affinity 5HT2a antagonist. Despite having a structural similarity to nonselective classical tricyclic antidepressant and antipsychotic agents, SAR and receptor modeling studies have suggested that AMDA and the classical tricyclic compounds interact differently with the 5HT2a receptors. It has also been suggested that the symmetrically folded aromatic geometry of the parent in the series, AMDA, is nearly optimal for 5HT2a receptor affinity.

The datasets contained in this thesis are broken into several groups according to the experimental data available and the feasibility of making direct comparison's between the atoms in the scaffold.

Group 1 AMDA with a 3 position substitution (Br, OH, C_6H_{13}, methoxy, CH_2CH_3Ph, $(CH_2)_4CH_3$)[29]

Group 2 AMDH with a 3 position substitution (Br, C_6H_{13}, methoxy, CH_2CH_3Ph, $(CH_2)_4CH_3$)[35]

Group 3 DOX with a 3 position substitution (Br, C_6H_{13}, methoxy, CH_2CH_3Ph, $(CH_2)_4CH_3$)[29]

Group 4 Cyproheptadiene and a related AMDA structure with a center ring substituent (X = CH_2, CH=CH, S, O, CH_2CH_2, H H)[36]

All geometries were optimized using Restricted Hartree-Fock (RHF) method with the DZV(2d,p) basis set. For the screening charge density calculations, the BP86/KTZP(2d,p) level of theory was used. The descriptor calculations were performed using the custom software described in Chapter 8. The following three sections describe the published experimental research associated with the Group 1,2,3 and 4 data-sets described above.

Group 1,2,3: Relation between AMDA and DOX agonists.

A 2003 study by Peddi et al[35] examined the effects of 3-position substitution of AMDA and compared the results to a parallel series of DOB-like 1-(2,5-dimethoxyphenyl)-2-aminopropanes substituted at the 4-position and 4-substituted 1-(2,5-dimethoxyphenyl)-2-aminopropane AMDH. The SAR data, results of receptor mutagenesis, and computer modeling of potential ligandreceptor binding modes for a variety of 5-HT2A agents suggests that ligands can bind in either of two overlapping sites: Site 1 and Site 2 as seen in Figure 7.13. All data suggest that DOB analogues can bind in either an agonist (interacting with TM3, TM5, and TM6; Site 1) or an antagonist mode (interacting with TM3, TM6 and TM7; Site 2) depending on the nature of the substituent. AMDA and derivatives are thought to share a binding mode with the antagonist

Figure 7.12: Parallel substitutions of AMDA and DOB-like and AMDH compounds

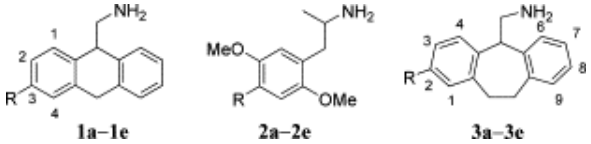
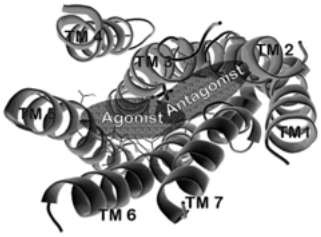


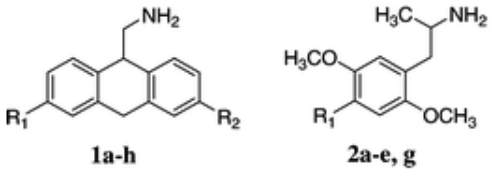
Figure 7.13: The 5HT2A antagonist and agonist binding pockets.



phenylethylamines regardless of the nature of the substituent.[35] AMDH was also able to be docked into the antagonist site in a manner similar.

In 2008 Runyon et al[29] performed an automated ligand docking and molecular dynamics study which suggested that all of the AMDA derivatives, the parent of which is a 5-HT2A antagonist, bind in a fashion analogous to that for the sterically demanding antagonist DOB-like compounds. The results were interpreted within the context of 5-HT2A receptor models that suggest that members of the DOB-like series can bind to the receptor in two distinct modes that correlate with the compounds functional activity.

Figure 7.14: Parallel substitutions of AMDA and DOB-like compounds

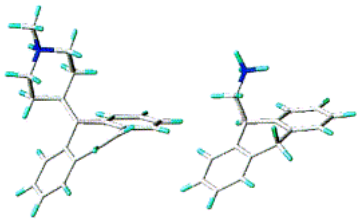


Group 4:Cyproheptadiene and AMDA Binding Mode

In 2001 Westkaemper et al[36] compared the serotonin 5-HT2A receptor affinities of a parallel series of structural analogues of AMDA and a structurally similar prototypical tricyclic amine cyproheptadine. The data suggests that the two agents bind to the receptor in different fashions. Examination of ligandreceptor model complexes supports the experimental data and suggests a potential origin for the differences in binding modes.

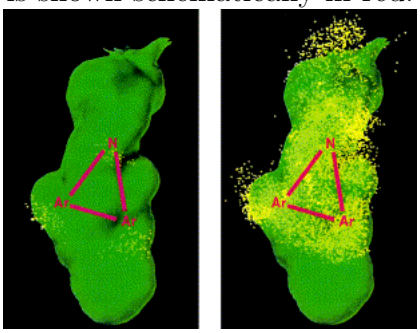
Computational simulations of the docking of cyproheptadine and AMDA to a 5-HT2A model were carried out in an attempt to identify potential similarities or differences in the modes of binding of the two ligands. Given the relative featurelessness of both structures, the only likely ligandreceptor interaction is that between the ammonium ions and Asp-155 (the highly

Figure 7.15: 3D image of AMDA and Cyproheptadiene



conserved residue mentioned previously 7.10). Manual docking followed by minimization and dynamics simulations produced results that were highly dependent on the starting configurations of the complexes. Since there are numerous potential starting configurations for complexes with either cyproheptadine or AMDA, these procedures are susceptible to an unacceptable level of operator bias. 7.16

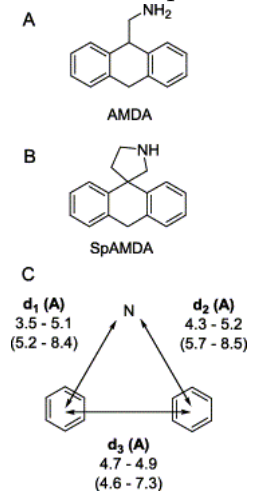
Figure 7.16: Plot of points (yellow) corresponding to N, C3, and C8 atoms for 86 sterically allowed conformers of cyproheptadine (left) and N, C3, and C7 for 4060 conformers generated from AMDA superimposed on a Connolly channel plot (green). The starting configurations of the ligands is shown schematically in red.



Group1, spAMDA derivatives: Synthesis of AMDA-like Molecules with Greater Selectivity

Shortly after AMDA was discovered to have such high affinity and selectivity a related molecule with even higher affinity was synthesized along with ring altered derivatives. spiro[9,10-dihydroanthracene]-9,3-pyrrolidine (spAMDA). These molecules seem to have the same basic pharmacophore but with different distances between the pharmacophoric groups. spAMDA plus it's derivatives that demonstrated reasonable affinity for 5HT2a are included in the Group 1 test set.[37]

Figure 7.17: AMDA pharmacophore



7.3 Method and Results: Agonist/Antagonist Discrimination Using the Screening Charge Density as a Fingerprint

Agonists and Antagonists have different but overlapping binding sites in the 5HT2a receptor. The presence of two distinct binding sites for GPCRs has been noted in the literature.[29] Consideration of ligand and receptor mutagenesis data led to the provisional conclusion of labeling these as an agonist site and an antagonist site. A comparison of the effects of a parallel series of aromatic substituents based on the tricyclic 5-HT2A antagonist AMDA suggests that the AMDA series may bind in a fashion similar to that of antagonist phenylalkylamines with bulky aromatic substituents.[35] The results were interpreted within the context of 5-HT2a receptor models that suggest that members of the DOB-like series can bind to the receptor in two distinct modes that correlate with the compounds functional activity. Automated ligand docking and molecular dynamics suggest that all of the AMDA derivatives, the parent of which is a 5-HT2A antagonist, bind in a fashion analogous to that for the sterically demanding antagonist DOB-like compounds. The determination of functional activity for 5HT2a (and the rest of the GPCR family) is experimentally challenging so computational tools that could add insight to what the determining factors could add valuable insights to what is currently known about the mechanism.

The details of the theory behind the charge density fingerprint was outlined in Chapter 2. In Chapter 4, it was shown in the analysis of the mono-substituted benzene set of molecules that the charge density histogram could represent the polarity profile and indicate the overall nucleophilic vs electrophilic nature of the molecules. Previous work has shown that the lipophilicity of 5HT2a ligands could be used as a general gauge of functional activity[30] so it was thought that the charge density profile might be of use in agonist vs. antagonist discrimination tasks. Given the success of neural network algorithms at pattern recognition tasks with histograms in images,[38] the charge density histogram was used as a test-case for the neural network infrastructure developed for this research. Neural networks can also be used as a modeling tool, much like linear regression to predict some output variable. This application of neural nets was also tried initially but was not successful at predicting Ki. This is not entirely surprising as the complexities of GPCR's make it notoriously difficult to predict Ki. Here, a supervised neural net architecture called the Multilayer Perceptron (MLP) will be applied to several discrimination tasks.

The basic approach for the analysis of the entire data-set (Groups 1-4, as well as groups of functionally active antipsychotics, antidepressents 102 cases) for classification of antagonist/agonist behavior using:

- MLP Neural Networks
- Logistic Regression

- Principle Component Analysis

Principle Component Analysis (PCA) and MLP Neural Networks were introduced in Chapter 3. Logistic regression is a variation of ordinary regression which is used when the dependent variable can only take two values. Unlike ordinary linear regression, logistic regression does not assume that the relationship between the independent variables and the dependent variable is a linear one. Logistic regression is a statistical method that allows group membership to be predicted from predictor variables, regardless of whether the predictor variables are continuous, discrete or a combination of both. It is an appropriate method to use when the dependent variable is expected to have a nonlinear relationship with one or more of the independent variables.

Neural Network classification of agonist vs antagonist of 102 cases

A schematic of the MLP Neural Network is shown in Figure 7.18. The set of boxes on the right, which represent the "bins" of the histogram. Each bin represents a charge density interval. The actual NN used had 30 bins (omitted for clarity in the figure). The middle layer of boxes represent the "hidden weights". The number of hidden weights is determined algorithmically for maximum performance. The last layer of boxes are the output neurons with the target classifications of agonist/antagonist. The neural net algorithm will optimize the weights based on the training dataset until the error is below some tolerance. After that, the network is tested with samples it has not seen before.

A representative sampling of the charge density profiles for the entire data-set is shown in Figure 7.19. These histograms are quite different from each other and the molecules represent many different structural and therapeutic classes. It is these differences that was hoped would make the sigma profiles a good candidate for a neural net pattern recognition algorithm.

The algorithm automatically partitions the data-set into randomly assigned training, testing and holdout samples. Both the training and testing samples are used internally by the neural net algorithm during the training procedure and the holdout samples are the true "test" cases that are reserved to assess the performance of the network. The neural network performed very well at the agonist vs antagonist discrimination task. Given that this particular data-set represented a large number of structural motifs and scaffolds it was somewhat surprising that the MLP neural net could perform so well on the full data-set. The intention with this analysis was to start with the entire data-set assuming the neural net would perform poorly and then reduce the data-sets into groups to improve the performance. Nevertheless, with this positive result in hand, the natural next step was to determine precisely what details in the charge density profiles were significant to the neural nets ability to classify these molecules.

Figure 7.18: Diagram of MLP Neural Network

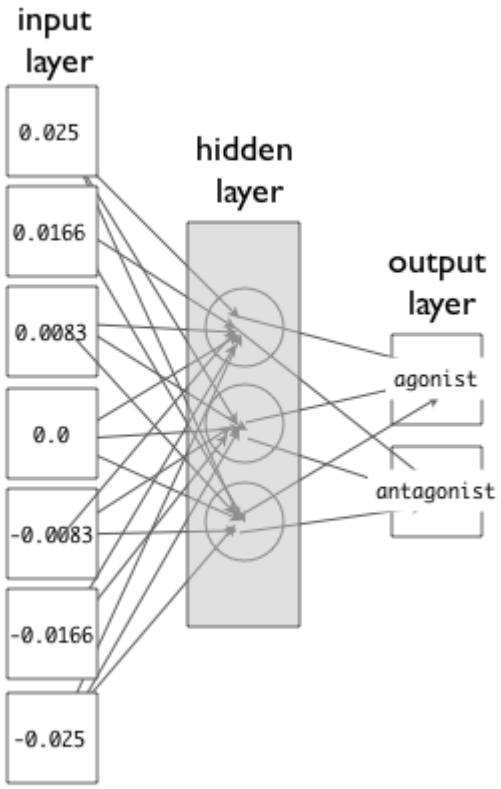


Figure 7.19: Charge Density Screening Histograms of All Agonists and Antagonists

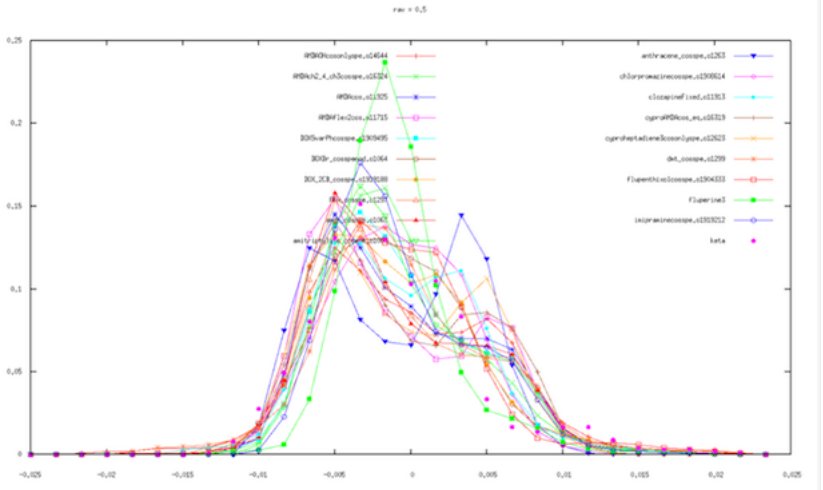


Table 7.1: Classification Results for Agonists and Antagonists

Classification

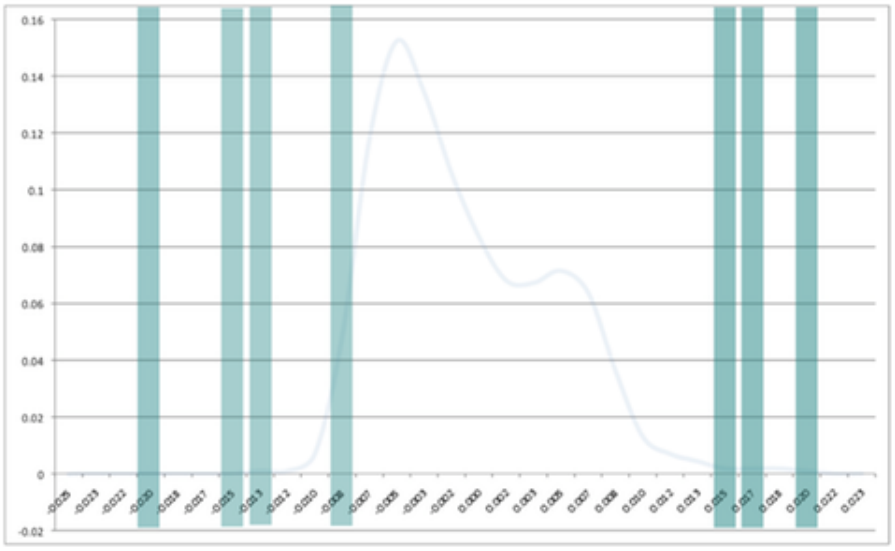
Sample	Observed	Predicted		
		agonist	antagonist	Percent Correct
Training	agonist	29	3	90.6%
	antagonist	4	39	90.7%
	Overall Percent	44.0%	56.0%	90.7%
Testing	agonist	7	1	87.5%
	antagonist	2	10	83.3%
	Overall Percent	45.0%	55.0%	85.0%
Holdout	agonist	4	0	100.0%
	antagonist	0	4	100.0%
	Overall Percent	50.0%	50.0%	100.0%

Dependent Variable: class

To get an idea of which charge density bins in the histogram contributed significantly to the success of the neural net the independent variable importance chart output was inspected (Figure 7.20). The variable importance projections are performed by using a procedure called sensitivity analysis which measures the effect that a change in the input variable has on the output variable. Sensitivity analysis is the study of how the variation (uncertainty) in the output of a mathematical model can be apportioned, qualitatively or quantitatively, to different sources of variation in the input of the model. Based on the results of the sensitivity analysis the user may eliminate redundant or irrelevant input variables. Reducing the number of input variables should provide better classification results.

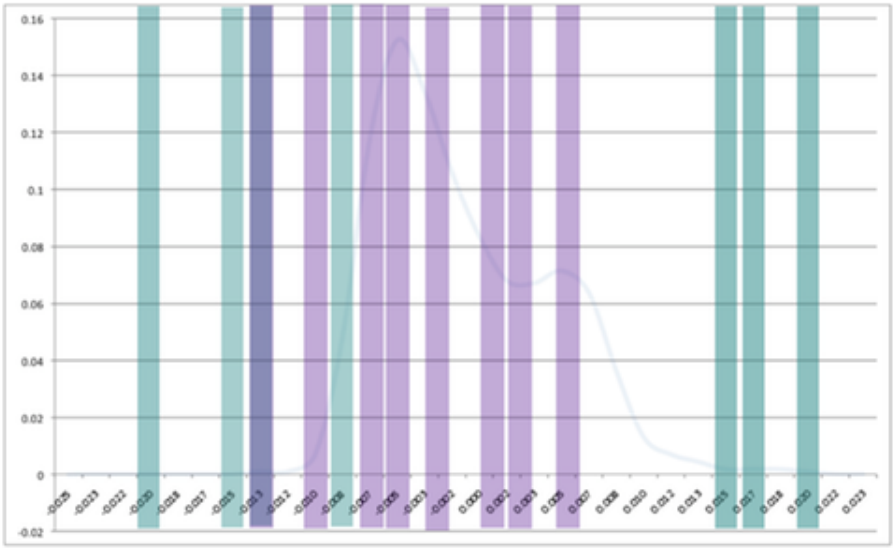
The peaks that the algorithm selected as "important" are shown in Figure 7.20. This particular plot is just an example (done in EXCEL) of a agonist for demonstration purposes. As was introduced in chapter 3, these outer areas of the histogram correspond to the polar regions of the molecule. Since the 5HT2a antagonists on average are more non-polar than the agonists it is not surprising that the largest differences are in these two regions.

Figure 7.20: Important Variables as Predicted by the Neural Net



To get an idea of the overall variance and the intercorrelations between the bins in the sigma profile the correlation matrix was examined and all the peaks with correlations < 0.95 and stddev > 0.01 were identified . This is shown in Figure 7.2 with the UFS selected variables in red and the NN results in green for comparison. Closer examination of the UFS criteria for elimination shows that the histogram bins were removed because of low standard deviation rather than multiple high correlations. UFS is just a data reduction technique applied only to the independent variables. It does not not guarantee the variables selected will be good predictors. By the same token, the variables at the extreme which were good predictors, and not selected by UFS may be highly correlated because of some common cause which also causes them to be strongly related to agonist, antagonist.

Table 7.2: stdev > 0.01 (purple) vs NN variable (green) selection



As an additional test of classification, a logistic regression was performed on the full data-set. Using all the variables (the Enter method) resulted in an overfit model so stepwise regression was used instead. The logistic regression also performed well at classifying the data-set into agonists/antagonists. (See Figure 7.3) with the final model (Step 4) correctly classifying 92.2 % of the molecules with an R^2=0.72 (Table 7.4. Stepwise logistic regression is similar to stepwise linear regression in that the variables are assessed one by one to improve the model performance so "Step 4" is the only relevant part of the table for analysis purposes.

Table 7.3: Logistic Regression Classification of Agonist/Antagonist

Classification Table[a]

Observed			Predicted		
			class		
			agonist	antagoni	Percentage Correct
Step 1	class	agonist	23	21	52.3
		antagoni	9	50	84.7
	Overall Percentage				70.9
Step 2	class	agonist	36	8	81.8
		antagoni	8	51	86.4
	Overall Percentage				84.5
Step 3	class	agonist	38	6	86.4
		antagoni	7	52	88.1
	Overall Percentage				87.4
Step 4	class	agonist	40	4	90.9
		antagoni	4	55	93.2
	Overall Percentage				92.2

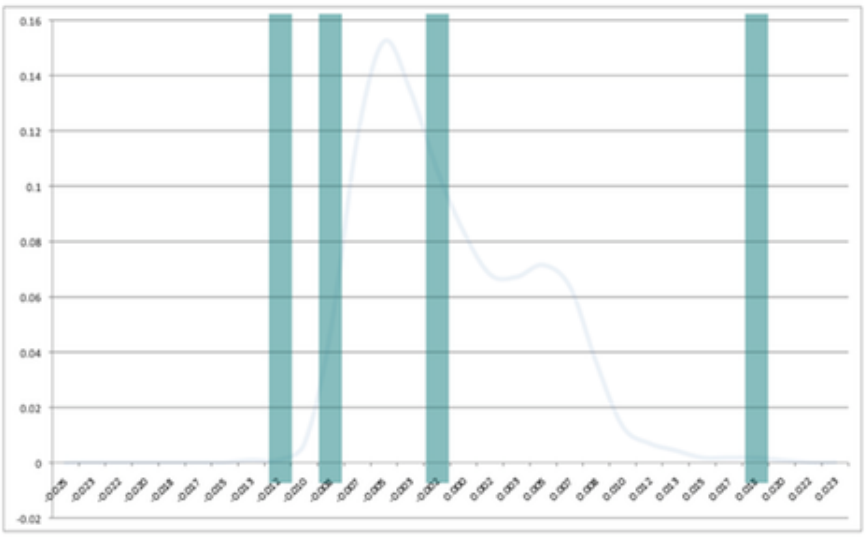
a. The cut value is .500

Table 7.4: Logistic Regression Model Summary of Agonist/Antagonist

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	112.838[a]	.247	.331
2	82.305[b]	.438	.588
3	68.264[b]	.509	.683
4	61.795[c]	.539	.723

Table 7.5: Logistic Regression Variable Selection



Only 4 variables were needed to classify the model (see Step 4 in Table 7.5) These are displayed graphically in Figure 7.5 with the important variables from the neural net analysis again superimposed for reference. It is interesting to note that the variables left in the equation after the logistic regression procedure are very similar to the ones selected by screening the correlation matrix (for corr < 0.95 and stdev > 0.01) and also completely different from the Neural Net. So it would seem that two different methods can yield good classification results but using differing portions of the charge density histogram. To further investigate, Principal Component Analysis was performed on the system to see if the overall dimension could be reduced into some key factors.

Principle Component Analysis of agonist vs antagonist of 102 cases

Given the differing results of the MLP neural net and the logistic regression, a principle component analysis was also performed to determine what regions of the charge density profile might be significant. The first four components make up for 74% of the total variance (Table 7.6). Some interesting patterns emerge when looking at the coefficients of the component matrix (Table 7.7). The significant values (coefficients > 0.6) are circled. Components 1,2 and 3 are displayed graphically in Figure 7.21 .The 1st component is comprised of the region that corresponds to positive charge density, the electrophilic "h-bond donor region" while the 2nd component covers the non-polar region within the "h-bond threshold" between $\pm$ 0.005 e/nm^2 and extends slightly into the nucleophilic, negative charge region. The third component is solely in the non-polar region of charge density.

PCA is not a procedure for classification of dependent variables. It reduces dimensionality in the independent variables, but the principle components/factors do not necessarily have any relationship to the dependent variable (agonist, antagonist). On the other hand, the structure

Table 7.6: PCA for charge density histogram. Total Variance Explained

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	7.156	26.504	26.504
2	5.926	21.948	48.453
3	3.658	13.550	62.003
4	3.245	12.017	74.020
5	2.261	8.375	82.395
6	1.351	5.005	87.400
7	1.049	3.885	91.284

Extraction Method: Principal Component Analysis.

Table 7.7: PCA for charge density histogram. Component Matrix

Component Matrix[a]

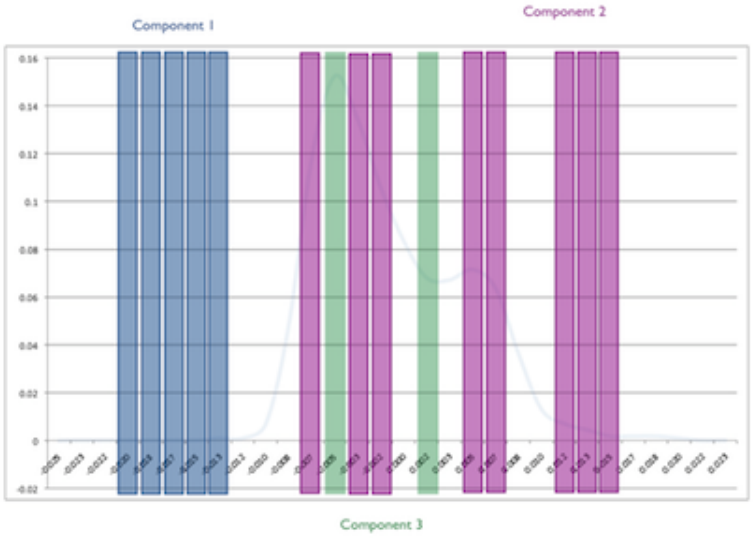
	Component			
	1	2	3	4
@0.02_A	.712	.214	-.236	-.017
@0.01833_A	.784	.224	-.315	-.140
@0.01667_A	.834	.200	-.338	-.177
@0.015_A	.869	.189	-.339	-.206
@0.01333_A	.888	.173	-.231	-.142
@0.01167_A	.793	.124	-.168	.056
@0.01_A	.568	-.243	-.243	.396
@0.00833_A	.261	-.714	.027	.522
@0.00667_A	.040	-.692	.509	.417
@0.005_A	-.218	-.094	.876	.157
@0.00333_A	-.563	.679	.250	-.285
@0.00167_A	-.586	.679	-.193	-.355
@0	-.599	.604	-.404	-.123
@0.00167	-.390	.266	.675	.481
@0.00333	-.006	-.440	-.599	.570
@0.005	.357	-.840	-.125	-.114
@0.00667	.458	-.629	.186	-.546
@0.00833	.441	-.302	.384	-.681
@0.01	.609	.266	.401	-.366
@0.01167	.496	.577	.206	.094
@0.01333	.364	.628	.176	.211
@0.015	.268	.616	.182	.247
@0.01667	.325	.583	.232	.430
@0.01833	.379	.396	.382	.480
@0.02	.372	.289	.400	.355
@0.02167	.318	.301	.357	.335

Extraction Method: Principal Component Analysis.
a. 6 components extracted.

of the Principal Components may provide a basis for classification of the independent variables. One way to check if the variables do have predictive ability, is to save the PCA Scores and then use these scores as variables for neural net or logistic regression. The saved scores are the data in terms of principle components.

The first entry for the saved scores, "FAC-1" is the coordinate of molecule 1 along the first principle component, and so on. It is the dot product of the 33 variables (corresponding to the 33 bins in the charge density histogram) in row one with the first principle component. The second row entry under "FAC-1" is the dot product of the 2nd molecule (second row of data) with the first principle component. So, this means the 103 molecule X 33 charge density bins data matrix multiplied by the matrix formed by first 5 principle components (33 molecules X 5 principle components) scores will be the 103 molecule X 5 principle data matrix in terms of the new variables (factors one through five). These results can be very hard to interpret since the

Figure 7.21: Visual Representation of Principle Components



scores represent a linear combination of all the variables unless there seems to be some latent physical meaning that can be ascribed to the components.

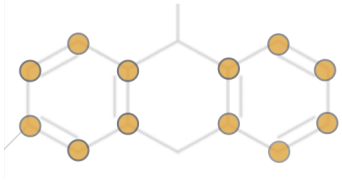
The PCA scores were used as variables in both the MLP neural net and logistic regression tasks to determine if these components had a predictive relationship with the output variable. For the Neural Net, FAC-1, FAC-2, and FAC-4 gave very good results (84.6% correct classifications) and FAC-2 alone gives acceptable results at 76.4% correct. For the Logistic Regression, the scores for FAC-1, FAC-2, and FAC-5 predicted 78.6% of the compounds correctly. Factors FAC-1 and FAC-2 gave an acceptable result (70.9%) but not quite as good as the Neural Net. Some components do great at classifying one variable while doing very poorly at classifying the other. For example, FAC-2 classified only 9% of the agonists correctly while classifying 74 % of the antagonists correctly.

7.4 Method and Results: Localized Property Investigation of AMDA

While 2D descriptors, such as the screening charge density histogram, have historically been successfully applied toward many areas of drug design, most notably, high throughput screening; they do not adequately represent the necessary detail to determine mechanistic details. Localized descriptors, on the other hand, can provide direct insight into the nature of the chemical interactions. For example, in chapter 3, the electron density differences were successful at revealing the ortho-meta-para directing nature of various substituents in mono-substituted benzene. In this chapter, the same localized descriptors introduced in the exploration of the set of mono-substituted benzene will be applied to the AMDA datasets. The details of the software developed for the parsing, extraction and analysis of this data is described in Chapter 8. The

QM descriptors were calculated at each atom center, except for the electron density density at 0.7 Bohr.

Figure 7.22: AMDA Scaffold for Localized Property Analysis



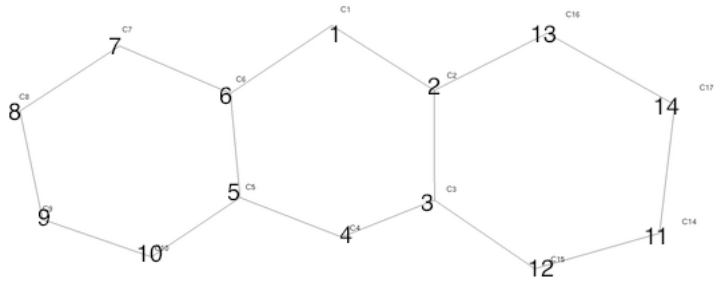
As described previously, the AMDA scaffold is a good choice as a representative tricyclic structure because of its selectivity for 5HT2a as well as its relative rigidity, which reduces the number of degrees of freedom substantially. The goal of this analysis was to explore the QM properties specifically on the core structure as shown in Figure 7.26. The focus of this analysis was to gauge the relative differences of this core structure rather than the direct interactions of the ligands with the receptor. A study of the direct interactions would be better suited to a docking approach, and given the uncertainties in the receptor structure of 5HT2a, the approach used here could provide valuable insights into factors governing the specificity of 5HT2a ligands. Additionally, the antagonist binding site of 5HT2a is lined with hydrophobic and aromatic residues, therefore an investigation of the small scale non-covalent interactions that contribute to binding is highly appropriate.

The AMDA test set was partitioned into the same subgroups described previously in section 2.3. The statistical analyses were done initially on each group separately for ease in interpretation of the results since each group is chemically similar and the experimental endpoint (binding affinity) was determined under the same conditions. After the groups were analyzed separately, and the most appropriate descriptor chosen, the groups were combined together. One of the challenges in this process was the small sample size of each group. For example, group 1 contains only 8 molecule cases, but there are 84 descriptors calculated per molecule. Clearly, the data must be preprocessed to avoid over-fitting and random correlations. small data-sets for a given emphtrusted experimental endpoint is typical of 5HT2a ligands. For example, many of the binding affinity studies cited in the Psychoactive Drug Screening Database[39] were performed using very different experimental conditions including: different radio-isotopes, receptors from different species, receptors for cloned emphin vitro studies vs in vivo, or receptors from different part of the brain. Data-sets from studies like these are not directly comparable.

Dataset: AMDA Group 1 (8 molecule x 56 descriptor dataset)

The initial data-set consisted of 6 descriptors: electrophilic/nucleophilic approximate superdelocalizability electrophilic/nucleophilic superdelocalizability on the scaffold shown in Figure 7.26. The frontier orbital densities were not used since the comparison was being done across molecules. There are 14 atoms in the AMDA scaffold under consideration (Figure 7.26). The

Figure 7.23: AMDA Scaffold fo Localized Property Analysis



first test set of molecules was the Group 1 described previously; AMDA with a 3 position substitution (Br, OH, C_6H_{13}, methoxy, CH_2CH_3Ph, $(CH_2)_4CH_3$).[29] This first analysis aimed to determine which descriptors on which atom were important to activity.

It was expected that some subset of descriptors on particular atoms would be the best descriptors of affinity. The full data-set is not ideal since there are many more variables than samples (56 variables with only 8 molecule cases). Many of the variables (e.g. approximate superdelocalizability, frontier orbital density and superdelocalizability) contain information that has a very similar theoretical basis, so perhaps one of them would do an adequate job on its own.

Analysis of the Transposed (56 descriptor x 8 molecule) Data-Set

The goal of the analysis of the transposed matrix was to determine if the scaffold could be reduced to a representative set of atoms rather than all 14. Ideally, a pharmacophore-like representation would be deduced. This would allow the combination of various group at later stages in the analysis. To explore the relationships between the molecules, the transpose of the original 8 (cases) x 56(descriptor variables) matrix was analyzed. Treating the molecules as variables makes the statistical problem simpler, as now the analysis can be done on a 8 x 8 matrix. This does not replace an analysis of the original 8 x 56 matrix, as it is needed to assess the relationships between atoms; rather this method will allow relationships amongst the molecules to become clearer. The transposed data set does not account for the correlations with the dependent variable (affinity) but only interrelationships.

The correlation matrix of the large descriptor set showed that all of the molecules were found to be highly correlated with each other, with several being perfectly correlated with each other. A small number of highly correlated observations can lead to an extremely unstable predictive model. One option for dealing with this problem could be to eliminate one or more of these molecules from the data. However, since there are highly correlated observations with significantly different values of the dependent variable (affinity) there is a problem of determining

which variables to throw out and which to keep. Another option is to combine or two or more of the cases (molecules) and use their average values. At the extreme, all eight cases could be averaged and replace them by an average case. This would be the crudest possible method and would essentially give us a single average data point which would not allow further mathematical modeling. Another option is to use PCA to try to obtain an estimator for the set of independent variables which is simpler than using all of the variables, and which accounts for much of the variance within the variables.

Figure 7.24: Transpose of the Correlation Matrix for AMDA Group 1

	AMDABr	AMDAOH	AMDA	AMDAc6h13	AMDAch2_3_ph	AMDAch2_4_ch3	AMDAmethoxy	spAMDA
AMDABr	1.000	.978	.971	.978	.978	.979	.979	.964
AMDAOH	.978	1.000	.997	.999	.998	.999	1.000	.989
AMDA	.971	.997	1.000	.991	.991	.993	.994	.994
AMDAc6h13	.978	.999	.991	1.000	1.000	1.000	1.000	.982
AMDAch2_3_ph	.978	.998	.991	1.000	1.000	1.000	.999	.981
AMDAch2_4_ch3	.979	.999	.993	1.000	1.000	1.000	1.000	.984
AMDAmethoxy	.979	1.000	.994	1.000	.999	1.000	1.000	.986
spAMDA	.964	.989	.994	.982	.981	.984	.986	1.000

Analysis of PCA Total Variance Explained

99% of the variance is explained by the first component, which implies that the eight molecules could be represented by one component. This is as should be expected from such highly correlated variables. The amount of variance explained by the components is equally distributed across all the components. One of the problems with PCA is that the interpretation of the components and how they relate to the original variables is not straightforward. Furthermore, if only one component is used, the data set is being reduced to one representative case; like using the average, just one case is inadequate for a predictive model. Table 7.4 shows a portion of the transposed data matrix. Notice that the column FAC1_1 which represents the PCA scores, and the Zavg which represents the standardized averages are very similar. This indicates that the variable reduction for the molecule provided by the PCA is in face not much of an improvement over just using the average.

Table 7.8: Principle Component Analysis with Scores

molname	AMDA	spAMDA	AMDABr	AMD...	AMDA_c 6h13	AMDA_metho xy	AMDACh2 _3_ph	AMDAc h2_4_c h3	FAC1_1	avg	Zavg
atom_0gamess.nuc_super	-6.67...	-6.1659...	-5.525...	-4.39...	-4.80...	-445.10409	-4.807...	-4.4...	-2.98148	-454.10	-3.04010
atom_0gamess.density	1.197...	1.19601E...	1.1961...	1.196...	1.196...	119.61212	1.1962...	1.19...	1.16708	119.63	1.06821
atom_0gamess.elec_super	-1.48...	-1.3236...	-1.376...	-1.40...	-1.37...	-14.32316	-1.446...	-1.4...	.07866	-14.12	.11053
atom_0gamess.approx_elec_s...	-1.84...	-.00806	-.06118	-6.24...	-5.65...	-.06932	-.06089	-6.2...	.19254	-.07	.21110
atom_0gamess.nucFODens	.06646	.04927	.00468	.00893	.07284	.00518	.00768	4.23...	.19327	.03	.21180
atom_0gamess.elecFODens	.05424	.00257	.01896	.01891	.01739	.02064	.01794	1.83...	.19325	.02	.21176
atom_0gamess.approx_nuc_su...	-6.08...	-.42773	-.06889	-1.20...	-6.04...	-.06565	-.07769	-5.1...	.19136	-.25	.20979
atom_1gamess.nuc_super	-9.00...	-7.4799...	-8.094...	-6.44...	-6.96...	-649.21329	-6.988...	-6.4...	-4.24604	-633.11	-4.32195
atom_1gamess.density	1.198...	1.19862E...	1.1984...	1.198...	1.198...	119.84670	1.1986...	1.19...	1.16896	119.85	1.06986
atom_1gamess.elec_super	-1.49...	-1.3389...	-1.372...	-1.40...	-1.38...	-14.28861	-1.443...	-1.4...	.07860	-14.13	.11041
atom_1gamess.approx_elec_s...	-7.10...	-.41254	-.02180	-2.64...	-3.88...	-.02363	-.02522	-2.5...	.19205	-.16	.21046
atom_1gamess.nucFODens	.08737	.08211	.01819	.02090	.05754	.01723	.02935	1.83...	.19339	.04	.21190
atom_1gamess.elecFODens	.20865	.13168	.00675	.00802	.01196	.00704	.00743	7.39...	.19341	.05	.21195

Conclusions from PCA of Transposed Matrix

Even though the number of atoms *could* be reduced to one based on PCA or using just the average; the result might be one molecule which has very high affinity, and so some small difference somewhere must have a huge impact. This argues for not collapsing the number of molecules. If the one with the largest affinity is the one that is perfectly correlated with another, then there must be some other factor explaining the affinity which is not in the model. From an analysis standpoint, if two molecules are perfectly correlated, then the PCA analysis will result in a singular matrix which could cause the computation to fail.

Universal Forward Selection

To determine which properties on which atoms might be significant in the original 8 molecule x 56 descriptor matrix, a data-screening procedure "Universal Forward Selection" was applied. Universal forward selection (UFS) is a data reduction algorithm[40] that selects from a data matrix a maximal linearly independent set of columns with a minimal amount of multiple correlation. It does this essentially by removing variables that have correlations greater than some value (the cut-off is typically R= 0.95) and a standard deviation less than some value (typically $\sigma < 0.01$). UFS will produce a reduced data set that contains no redundancy and a minimal amount of multicollinearity.

Since there are far too many variables, it was thought that this method could be applied to the entire dataset as a first step in the analysis. The initial UFS screen was performed on the original 8 x 56 data-set, but the results did not yield any meaningful trends. The UFS screening procedure can be used with or without an independent variable and the analysis was tried both ways. In both situations, the variables that were statistically selected did not appear to have any chemical meaning regarding atoms or properties. To solve this the data-sets were refined. The single 8 x 56 variable data-set was divided into 7 smaller data-sets according to property (electrophilic/nucleophilic Frontier Orbital Density, electrophilic/nucleophilic approximate superdelocalizability electrophilic/nucleophilic superdelocalizability and the electron density at the nuclei) making 7, 8 x 14 data-sets. The UFS results for these data-sets also gave conflicting results with different atoms and different properties. In each case, it was not possible to find a common set of variables that are predictive across all the atoms. Without a scientific rationale for using the UFS variables, there is a danger of just fitting random noise. Since UFS is just meant to be a data-screening device these results were not entirely surprising, and more rigorous methods were explored for data-screening.

The UFS removes the variables with low standard deviations because they are almost constant and therefore can contribute very little to analysis. That is, all the molecules have the same or almost the same values for them since variation in x is required to explain variation in y. UFS only removed 5 of this type. Next, it removes variables which are highly correlated with each other. If the overall goal of the analysis was linear regression aimed at prediction, these

highly correlated variables are measuring almost the same thing, and so they are redundant. However, from a scientific or analytic point of view, exactly emphwhat might be a common influence on these correlated variables is important and emphwhy they are so correlated.

Stepwise Linear Regression

Since the number of variables is so large and the number of cases so small, over-fitting of the variables is a danger. Seven independent variables are all that would be needed to perfectly fit the model. Meaning, it would be possible to just generate random entries for the variables and still find variables that would be predictive. Care needs to be taken in analyzing the results from a linear regression in this situation.

Stepwise regression compares each variable with the dependent variable one by one, and then throws it out or brings it into the linear regression model at each step until it gets a good fit. While the stepwise method is better when over-fitting is a potential problem, the method will not select any variables if none of them are sufficiently correlated with the dependent variable Ki. The stepwise regression was repeated for all 7 of the reduced data-sets. The variables selected (approx_elec_super, elec_super and nuc_super) selected, were combined on the 3 atoms (6, 7, 8) and a linear regression (enter method) yielded a fairly good fit (R square = 0.961). Not too much can be read into this result however. Problems remain from an analytic point of view in that given the large number of variables, one could almost certainly just generate random entries for the variables and find 4 to six that would be predictive. 7 independent variables are all that would be needed to perfectly fit the model.

Atom by atom regression

A brute force atom by atom regression with the affinity values was performed to determine the correlations of each atom with affinity. Some atoms were much better than others, with an adjusted R^2, value that is greater than 0.900. More importantly, some of the atoms had an adjusted R squared value that was negative which means that they should be excluded from the model. Adjusted R squared values were used in this analysis because adjusted RR^2 values penalizes for complexity in the model. When the adjusted R^2, is much lower than the regular R_2, that is an indication that there are more variables than needed, and that a high regular R^2 is obtained just by the virtue of having so many variables. R^2 can be increased, even to point of perfect fit, just by adding more variables- even if their values are assigned at random. In this atom by atom analysis, since some of the adjusted R^2 were much lower that the regular R^2, if the R^2s were used as selection criteria, the result would end up picking some atoms that almost certainly would be not good ones, and would not be useful for predictions outside of this immediate data set.

Conclusions for Analysis of AMDA Group 1 (8 molecule x 56 descriptor dataset)

While these results on their own do not establish that these atoms would be significant to the prediction of Ki, they do indicate that certain atoms would be better choices than others for the next step in the analysis and certainly that some atoms should be excluded from the model. It could be that the property variables just are not predictive, as evidenced by very low correlation with Ki, and any linear models that provide a good fit are just the result of chance. If that is the case, then the model would probably not be generalizable to other molecules. Still, it may be possible to determine some structure by examining the latent factors in the partial least squares regression, and find important variables.

Analysis of AMDA scaffold Group 1, Group 2 and Group 3 with electron density (38 molecule x 14 descriptor dataset)

The data set was expanded to include data-sets from Group2 and Group 3 making the data-set size 38 molecules total. The analysis was restricted to the carbon atoms in the AMDA scaffold as shown in Figure 7.26. The same statistical analysis methodology was applied as for the set above: correlation matrix of the transpose, UFS, correlation matrix, stepwise regression, PCA and PLS to determine which atoms are important to output value (the affinity, Ki). For these data sets, the electron density at 0.7 Bohr was used as the sole descriptor to minimize problems with interpretation.

Analysis of the Transposed (14 descriptor x 8 molecule) Data-Set

The transposed correlation matrix was examined for easier identification of multiple correlations between molecules. The same procedure as for the 8 x 56 dataset was repeated to determine if the molecule representations could be collapsed. Unlike the previous analysis, there were no perfectly correlated molecules that necessitated removal.

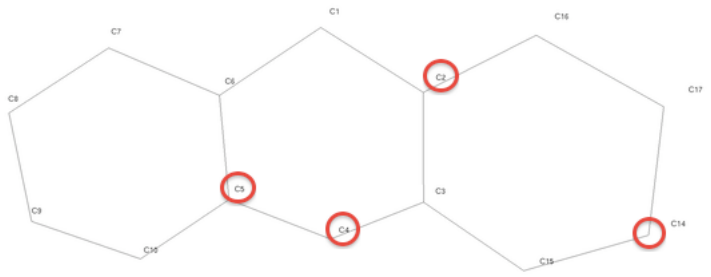
Analysis of PCA Total Variance Explained

Unlike the situation for the 8 x 56 dataset, excessively high correlations were not found between compounds, so reduction of the number of molecules was not desirable. Additionally, the PCA also did not add any insight into the structure of the independent variables.

Universal Forward Selection

The UFS procedure selected atoms 2,4,5 and 14. As was the case for the analysis of the complete data previously, the UFS variable selections did not have predictive power and were not used for variable selection.

Figure 7.25: Expanded Dataset, UFS Results for Electron Density at 0.7 Bohr



Stepwise Linear Regression

Linear Regression

As was the case for the 8 x 56 dataset, linear regression was used to give insights into which, if any, atoms contributed significantly to the observed affinity. Linear regression using all the variables yielded a model with a very low R^2 of 0.492 and an Adjusted R^2 of 0.168 so this model was not considered. The stepwise regression selected atoms 5 and 10 with a very low R^2 of 0.185.

PCA results

The principle component analysis of this dataset did not reveal any obvious latent structure in the data (Table 7.10) based on the atom groupings. It is interesting to note that there are indeed 3 distinct components which represent different groupings of atoms. Given the diverse nature of this diverse data-set (Group 1, 2 and 3) however, it is difficult to draw chemical conclusions based on these groupings.

Table 7.9: Expanded Dataset, PCA Total Variance Explained for Offset Electron Density

	Initial Eigenvalues		
Component	Total	% of Variance	Cumulative %
1	27.521	72.425	72.425
2	4.367	11.491	83.916
3	2.618	6.890	90.806
4	1.075	2.829	93.634
5	.880	2.317	95.951
6	.599	1.577	97.529
7	.360	.949	98.477
8	.227	.597	99.074
9	.182	.479	99.553
10	.087	.229	99.782
11	.038	.100	99.881
12	.031	.082	99.964
13	.014	.036	100.000

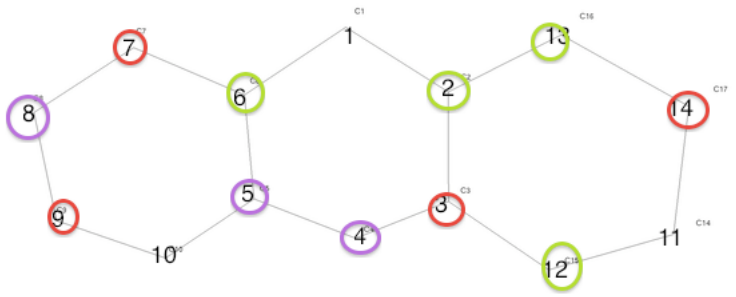
Extraction Method: Principal Component Analysis.

Table 7.10: Expanded Dataset, PCA Component Matrix for Offset Electron Density

	Component				
	1	2	3	4	5
atom1	-.141	.390	.395	-.056	-.595
atom2	-.529	.574	.128	.351	-.024
atom3	.813	.444	-.165	.087	.140
atom4	.163	-.075	.572	.596	.358
atom5	.559	-.159	.596	.077	.437
atom6	.548	.592	-.002	.321	-.114
atom7	.855	-.022	.021	.108	-.234
atom8	-.239	.454	.554	-.089	-.320
atom9	.647	.174	.367	-.259	-.145
atom10	-.405	-.561	.286	.287	-.099
atom11	-.288	.713	-.180	-.134	.405
atom12	-.313	.599	-.315	.605	-.024
atom13	-.101	.614	.249	-.603	.303
atom14	-.901	-.017	.334	-.056	.149

Extraction Method: Principal Component Analysis.
a. 5 components extracted.

Figure 7.26: Component 1 in Red, Component 2 in Green, Component 3 in purple



PLS results

Table 7.11 below has the latent factors listed by row. The cumulative Y variance is the percent of variance in the Y variable accounted for by the latent factors. Likewise, the cumulative X variance explains the variance in the X factors. Unlike the PLS analysis of the triptycene dimers in chapter 4, there is not one component that significantly explains the variance in x or y, and the R^2 values are very low indicating that this is not a predictive model in any case.

The coefficients given in the `Variable Importance in the Projection` reflect the relative importance of each X variable for each X factor in the prediction model. Since the Y-scores are predicted from the X-scores, the VIP coefficients represent the importance of each X variable in fitting both the X and Y scores. Often times, independent variables whose VIP coefficient is ¡ 0.8 and also have a small regression coefficient are removed from the model. Based on the VIP coefficients in the table below, atoms1, 2,3,6,8,9,12,13 and 14 could be candidates for removal, while atoms 5 and 10 have large VIP coefficients. Before deciding definitively which

Table 7.11: Expanded Dataset, PLS Proportion of Variance Explained for Electron Density at 0.7 Bohr

	Statistics				
Latent Factors	X Variance	Cum X Var	Y Var	Cum Y Var	Adj R-square
1	.201	.201	.229	.229	.207
2	.189	.390	.064	.293	.251
3	.130	.521	.035	.328	.267
4	.125	.646	.029	.357	.276
5	.042	.688	.057	.414	.319

variables could be removed the regression coefficients should be examined. The `Parameters` table (Table 7.12) gives the Regression parameter (coefficient) estimates. These are the regression coefficients and indicate the rate of change of the dependent variable per unit change in the independent variable (listed on the left). From this table variables that don't have much influence on changes can be identified. In this case, all the coefficients are very large so do not aid in selection criteria. The "Proportion of Variance Explained" indicate that atoms 5 and 10 were important and 7,11 and 12 were also important to a lesser degree. However given the overall low R^2 of the model these results should not be considered conclusive and no variables were selected as candidates for removal.

Table 7.12: Expanded Dataset, PLS Parameters for Electron Density at 0.7 Bohr

	Variables
Independent Variables	output
(Constant)	51881.834
atom1	-4063.544
atom2	70781.005
atom3	-76938.113
atom4	1374.899
atom5	-72399.222
atom6	70274.038
atom7	-12519.605
atom8	-113945.304
atom9	15992.568
atom10	-26457.764
atom11	84082.218
atom12	-82719.649
atom13	-81204.622
atom14	-1504.638

Conclusions for AMDA scaffold Group 1, Group 2 and Group 3 with electron density (38 molecule x 14 descriptor dataset)

The electron density at 0.7 Bohr was not highly correlated with affinity for any of the atoms; consequently the R^2 and adjusted R^2 are too low to be meaningful with ordinary linear regression models. Furthermore, the PLS models are also not satisfactorily predictive. It could be the case that a linear model might not be adequate for this data and non-linear methods such as Artificial

Table 7.13: Expanded Dataset, PLS Proportion of Variance Explained for Electron Density at 0.7 Bohr

	Latent Factors				
Variables	1	2	3	4	5
atom1	.654	.848	.813	.809	.752
atom2	.666	.659	.627	.724	.738
atom3	.168	.340	.691	.930	1.728
atom4	.875	.774	.926	1.075	1.018
atom5	1.863	1.690	1.600	1.536	1.461
atom6	.324	.305	.308	.591	.893
atom7	1.376	1.218	1.198	1.149	1.163
atom8	.575	.850	.822	.789	.901
atom9	1.140	1.008	.952	.978	.974
atom10	1.822	2.162	2.172	2.286	2.203
atom11	1.134	1.003	.974	.954	1.008
atom12	.701	.727	.894	.923	.967
atom13	.248	.364	.514	.527	.721
atom14	.281	.648	.635	.612	.577

Cumulative Variable Importance

Neural Networks (ANN) might be more appropriate.

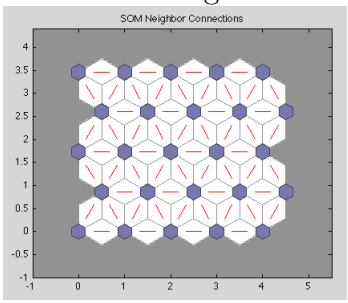
Self Organizing Map Results

As was introduced in Chapter 2, the Self-Organizing Map (SOM) is an established method for mapping data from a high-dimensional space to a lower dimensional space. In particular, when the original data is mapped onto a 2D plane, often called a `SOM map` relationships are sometimes easier to visualize. There are many data-reduction techniques, such as Principle Components Analysis (PCA) discussed previously; one of the advantages of using a SOM for data-reduction is that the topology is maintained; data points located close to each other in the original space are also close together on the SOM map. This is particularly advantageous for structural data since it is desirable to preserve the relative location of the data points. Much like principle

component analysis, the hope with unsupervised learning methods such as the self-organizing map (SOM) is that the method will highlight some latent correlations in the data-set.

Like the MLP Neural Net introduced in Chapter 4, the SOM consists of a set of neurons and connections between them, as shown in Figure 7.27. The blue hexagons represent the neurons and the red lines indicate the connections between neighboring neurons.

Figure 7.27: SOM Neighbor Connections

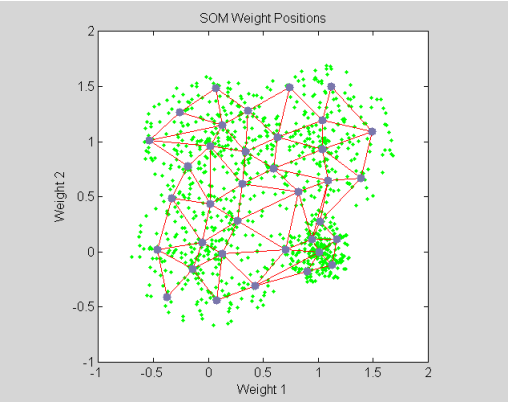


The way the neurons are connected in the SOM is different from the MLP in that they are connected to each other in such a way that when the SOM is first `trained` each neuron's weight vector becomes more similar to the input vectors. Unlike the MLP, there is no target output that the algorithm uses to adjust the weights. This is why the SOM is termed an "unsupervised" neural net. As the training proceeds, each neuron's weight vector becomes more similar to the input vectors. Figure 7.28 show 2 weights as a demonstration. The green-dots are the original data points and the gray dots are the neurons with the red lines representing the connections. The neurons are initially given small randomized (between $\pm$ 0.01) values and as training proceeds the weight vectors become similar to the input vectors and the neurons should be distributed somewhat evenly over the data set.

The data-set used here is the same 14 atom AMDA scaffold with electron density values at 0.7 Bohr from the plane of the rings used in the previous parts of the chapter, so each input pattern would represent the 14 values for the electron density. Each neuron actually has 14 weights corresponding to the 14 atoms. When the input space is high dimensional, all the weights cannot be visualized at the same time. For this analysis, a 5 x 5 neuron SOM was trained using all 35 of the molecules in the AMDA dataset used in the previous analyses in this chapter: described in Chapter 1: Group 1 (the AMDA scaffold and an 3 position substituent) , Group 2 (AMDH with a 3 position substituent) and Group 4 (a cyproheptadiene like AMDA structure with a center ring substituent) .

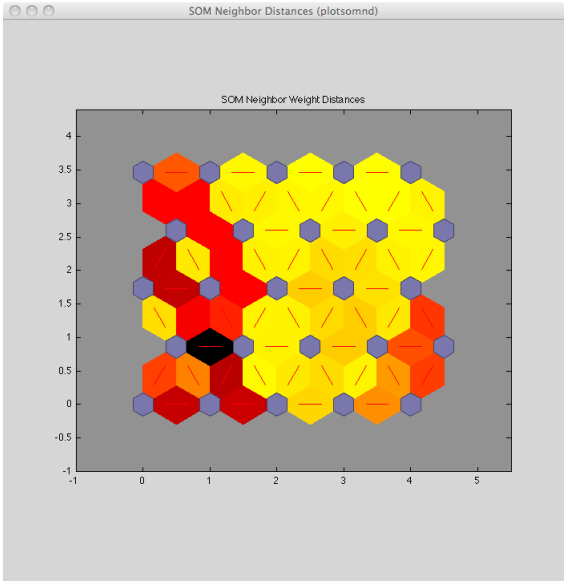
Since the input space is high dimensional (14 atoms) all the weights cannot be visualized at the same time. Instead, it is more intuitive to color code the distances between the neighboring neurons. The blue hexagons represent the neurons and the red lines indicate the connections between neighboring neurons. The coloring of each region containing the red line indicates the distance between each neuron. The darker colors represent larger distances while the lighter colors represent smaller distances. Notice that in Figure 7.29 the lower left corner contains

Figure 7.28: SOM Weight Positions



a grouping of darker colored segments. This indicates that the network has roughly clustered the data into two groups: the yellow area would represent input patterns with small weights (distances), while the darker areas represent input patterns with larger weights. The identity of these clusters must be attributed to the classifications manually in a manner described next.

Figure 7.29: SOM Neighbor Weight Distances



After training is complete, the neurons can be colored or labeled according to the class. In order to accomplish this, an input pattern is again presented to the network and the algorithm compares it to all the neurons in the output layer. The method that adjusts the relative weights that represent the connectivity between the neurons is a 'winner-takes-all' algorithm, such that one neuron vector will be selected which has weights most similar to the input pattern. This neuron will be the 'winner' and will become 'active' by firing a signal. The remaining neurons are inactive. Each pattern is assigned to exactly one neuron in this fashion. The data patterns assigned to a particular neuron are more similar to emphit's own neuron than any other and will form a cluster around that neuron. After the training process was complete, each molecule

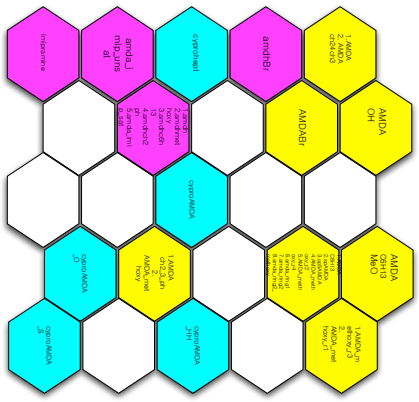
pattern was presented to the network individually and the neuron that was activated was labeled as shown in Figure 7.30. The colors represent the classifications which were done in this case by inspection. This is typically how the classifications are done. Yellow represents the AMDA-like scaffold of Group 1, purple is the imipramine-like AHDH scaffold and cyan is the cyproheptadiene like scaffold. The text on each individual neuron lists the test patterns that activated the neuron. This is not an important detail except to note that some of the neurons were activated by several test patterns.

Considering the small size of the training data-set the SOM did surprisingly well at separating the data into 3 classes based on the electron densities of the 14 atoms of the AMDA scaffold. The conditions of the analysis were not ideal, mainly there are not enough training samples. This is evidenced by the large number of in-active neurons which were not activated by any of the test patterns. The evidence of not enough training samples is demonstrated by the large number of inactive neurons during the initial training of the SOM.

Another factor that was non-ideal was the size of the overall SOM map. A map size of approximately 10 x 10 map would be closer to ideal, but there were not enough samples for a 10 x 10 map to be feasible for this analysis. This problem is also discussed in greater detail in chapter 7. Even with the obvious shortcomings in the prerequisites for a successful analysis, the results indicate that a SOM map could be a useful tool for classification of the data.

Visualizations such as this serve to assess how suitable the chosen molecular representation could be in a classification task. For this analysis, in addition to the electron densities, all the descriptors under-discussion in this thesis were evaluated in the manner just presented (nucleophilic/electrophilic frontier orbital densities, nucleophilic/electrophillic approximate superdelocalizabilites, nucleophilic/electrophillic superdelocalizabilites). Based on the initial distribution of the SOM neighbor connections Map (Figure 7.28) the electron density was the best candidate for the SOM classification task.

Figure 7.30: SOM Classification Map



Conclusions for Charge Density Histogram Investigation

The logistic regression and neural nets perform well at classifying the dataset into agonist and antagonists, but it is a little more difficult to determine what variables are most predictive, and why. The PCA results combined with the NN and logistic regression analysis give the best indication of what variables are most significant in the classification ability of the methods.

For this particular analysis, if it is assumed that component 1 corresponds to nucleophilic positive charge density, component 2 corresponds to electrophilic negative charge density and component 3 corresponds to nonpolar behavior, according to the regions of charge density these components occupy, then some tentative conclusions can be drawn from the PCA analysis after determining if the components indeed have any predictive behavior. Antagonists, while generally nonpolar, tend to have a right peak that is larger in surface area than the left, this corresponds to more surface area in the molecule having a positive charge density which for antagonists corresponds to the positive hydrogens. Agonists typically have one more pronounced peak also in the positive charge density area with a much smaller peak in the negative charge density area. Component 1 is in the region where both the antagonists and agonists have a peak (of differing height and slightly different position) from their positively charged hydrogens. Component 2 is in the region of the histogram that is indicative of the negative charge density of the aromatic face of the benzene rings which is greater for agonists than antagonists. In this way the PCA has helped to elucidate the significant portions of the charge density profile for agonist/antagonist behavior.

Conclusions for Localized Properties Investigation

After investigating the chemical scaffold of the Group 1 dataset using a variety of statistical methods, it could be concluded that, in principle, the the structure of the system could be simplified by eliminating redundant molecules, by using a single molecule represented by their average values, or by using principle components of the transposed data to represent the eight with a smaller set of one or two. However, this avenue was not taken so as to not lose resolution in using an average; and in using either the average or principle components to reduce the number to less than eight would not be able to take advantage of the information in the 8 Ki values.

Reducing the number of variables, however, is essential for a good model. The dimension of the data set is less than or equal to seven (since two molecules of the eight were perfectly correlated), so any model with seven or more variables is certain to be overfit. Preliminary findings in using stepwise regression indicate that the descriptor measurements on any atom could well fit the model. However, arbitrarily selecting one of the atoms may cause loss of valuable information carried in the others because even though the atom are similar they are not identical. The principal components capture the variation in the independent variables very well, but it was also found that regressing on the principle components does not give a good fit,

meaning that the PCA factors they would provide a poor basis for variable reduction.

PLS yielded the best results to keep the eight molecules in a model that is both parsimonious and predictive, and which captures all the information, in the data set. The PLS predicts Ki very well using only 5 components. A reasonable prediction can be obtained using only two or three PLS latent factors. The PLS procedure uses the scores (i.e.; latent variables) as the variables to predict Ki. The simplest predictive mode found was to use the average the scores. The average scores predict better than using the first four latent vectors and almost as well as using all five.

Turning to the second dataset (Groups 1,2,3) with the offset electron density as the predictor, elimination of redundant molecules was not required; and PCA failed to identify any meaningful relationships between them. All of the linear models used,including PLS,were found to be inadequate in predicting or explaining the dependent variable (output in dataset, represents affinity); this is not surprising since none of the independent variable in this dataset are highly correlated with the dependent variable (affinity). One possible explanation for the weak correlations is that in the analysis of the offset electron density, perhaps the three groups are different enough such that that a linear relationship cannot be found that spans them. Another possibility is that the results obtained with the analysis of group 1 was a result of having so many variables and so few cases that the relationships found were not meaningful or representative and that the lack of linear relationship found in the Groups 1,2,3 dataset is in fact the more representative. This question cannot be satisfactorily resolved through data analysis alone. Since the electron density is a continuous property, a more representative sampling of the electron density in the area around the molecule rather than just at the points directly offset from the nuclei at 0.7 Bohr would provide a better representation of the reactive space of the molecule. One of the major guiding principles for the work described in this thesis was to develop the infrastructure to link visualization at the molecular surface with analysis tools to analyze data scenarios such as this one. While the infrastructure is in place (described next) more development needs to be done to incorporate the statistical methods described into the overall data structures so that a quantity like the electron density can be calculated *and* analyzed on the molecular surface. Likewise, the preliminary SOM results indicate that the offset electron density can be used as criteria for classification by unsupervised neural nets, and the possibility of integrating the SOM as a secondary means of visual analysis could also be implemented in the analysis infrastructure.

Chapter 8

Computational Infrastructure and Methods

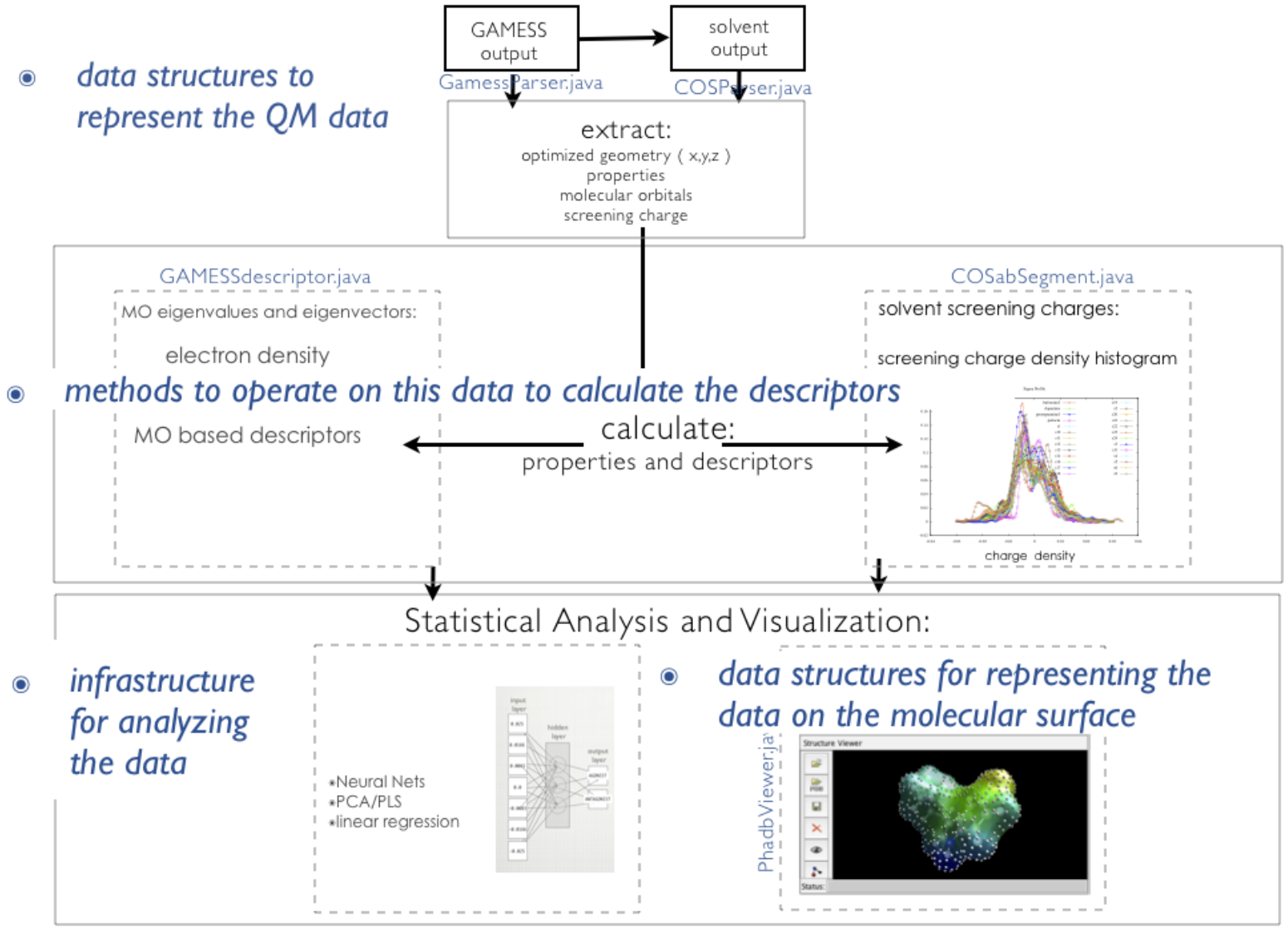
8.1 Overview

To carry out the studies described herein, tools had to be developed for the parsing, analysis and interpretation of the data. This effort falls into four main categories: (summarized in Figure 8.1): 1) data structures to represent the QM data , 2)methods to operate on the QM properties to calculate descriptors, 3) the development of the infrastructure for analyzing the data and 4) data structures for representing the data on the molecular surface. Chapter 1 and 2 detailed the theoretical background and basis for the QM properties and descriptor calculation, here the focus is on the computational infrastructure and implementation details themselves. All methods were written in Java.

The first step in the process (Figure 8.1) is to parse the GAMESS primary output file as well as the associated secondary ".cosmo" file which contains the output from the solvent code. Custom parsers were developed to parse the geometries, properties, molecular orbitals and screening charges from these output files. This data is stored in custom data structures that were designed to enable efficient calculation of descriptors in the next step: descriptor calculation. The bulk of the methods written for the descriptor calculation use the molecular orbital eigenvector and eigenvalue matrices to calculate descriptors based on the molecular orbitals or the electron density. Also, the screening charge density histogram is constructed from the segment and screening charge densities in the solvent output file. The descriptors are stored in the appropriate data structure for representation and analysis during the final phase of the process. The infrastructure for correlating the descriptors with the desired endpoint (affinity) involved the creation of a number of different utilities for coupling the structure and property data with already existing software and libraries. Custom methods for the Neural Network analysis were also created for analysis purposes. The data structures and algorithms to support the visual representation and analysis were developed using a Delaunay based triangulation. The surfacing effort includes both a custom coded implementation of the Convex Hull algorithm as well as a

implementation of the CGAL triangulation data structures and libraries.

Figure 8.1: Computational Infrastructure and Methods Overview



This development was guided not only be the scientific research goals but also by the requirements of two external projects: a QSAR workflow environment "Chemomentum" and a molecular graphics visualization project Sirius . The overall scope of the research in the Chemomentum project is outlined in Figure 8.2. Both of these efforts hinged upon utilizing quantum mechanically computed properties from the GAMESS software package. To avoid redundancy and maximize effort an open-source cheminformatics library the "Chemistry Development Kit" was used as a starting point in the parsing efforts. To effectively utilize this library the existing data structures needed to be supplemented to include concepts used in quantum mechanics such as molecular orbitals, eigenvectors and eigenvalues. This supplementation is represented schematically in Figure 8.3

To accommodate the requirements of the Chemomentum project, flexibility in the input and output formats were added to the workflow environment. Also, the individual java classes were extracted to be stand alone jar files so that they could operate independently. This included a modularization of the visual components used for Sirius such that the visual analysis could be executed in batch mode on a series of molecules at once rather than using a traditional graphical user interface to open molecules one at a time.

Sirius was selected as the framework for the molecular visualization efforts because of its extensibility in both the direction of macromolecular protein structures as well as small molecules. These met the needs of the perceived goals of the project for linking quantum mechanically derived data and experimental endpoints from biological receptors. The Baldridge group successfully utilized "Garnet" a java applet version of Sirius in a previous workflow effort "GEMSTONE" (a workflow environment from the San Diego Supercomputer Center) and it was hoped to further development of Sirius in this direction within the Chemomentum workflow project. Sirius is a component-based visualization system originally developed at San Diego Supercomputer Center by Sasha Buzko. Sirius provides tools for molecular modeling, drug discovery, protein structure analysis, as well as data mining and sequence-based work. It includes Structure Viewer (3D display), Sequence Viewer, and Structure Browser. These components are linked to allow simultaneous updates to the displayed data in response to changes, such as structure edits and appearance changes. Sirius provides a convenient interface for several third-party applications. These include Modeler (homology modeling of protein structures), Amber and CHARMM (molecular dynamics setup and trajectory visualization), POV-Ray (high-quality ray tracing for scene rendering), as well as structure and sequence alignment functionality. A dedicated data access component provides the ability to run BLAST and InterProScan searches in the Uniprot database.

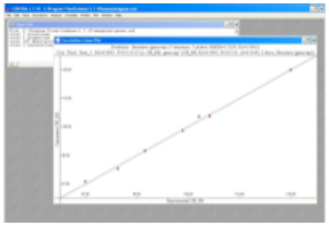
Figure 8.2: Chemomentum Project Goals

Overview

Application I:
"Chemomentum" :
traditional QSAR and high-throughput workflows on distributed computing resources.

Goal 1:

integration of QM data
data structures for more accurate descriptor calculation



GAMESS output

UzhGamessParser.java

extract:
(gas phase and solvated properties)

optimized geometry
properties
molecular orbitals
screening charges (sigma profile)

Phadbeta.java

calculate:
properties and descriptors using the improved CDK libraries

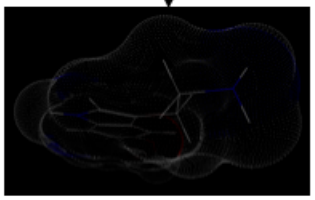
SOMNet.java

regression:
correlation between structure and activity using Neural Net algorithms.

SurfaceMap.java

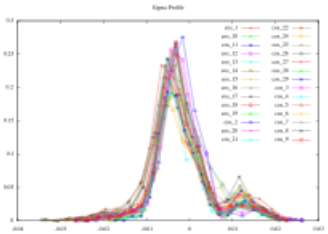
visualization/mapping:
properties mapped onto a surface

PhadbViewer.java



Application II:
Receptor surface model.
Modeling GPCR's and CNS drugs.

Goal 2:
Investigate Neural Net methods for pattern recognition/non-linear regression.



Goal 3:
QM Properties mapped on to hypothesis of receptor surface model.

4/11

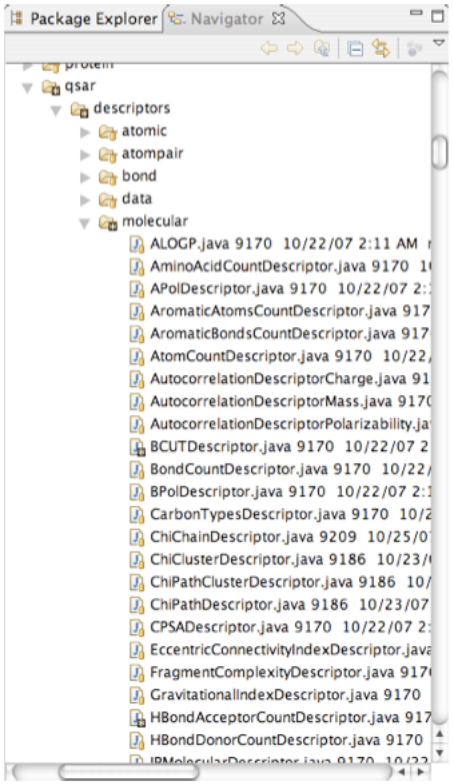
8.2 Cheminformatics and Visualization Tools and Libraries

The Chemistry Development Kit (CDK) is a freely available open-source Java library for Cheminformatics and Bioinformatics. The CDK itself is designed to be a library rather than a standalone program.[41] It provides methods and data structures for many of the common tasks in molecular informatics. For the work described in this thesis, functions were added to support the extraction of quantum mechanical data and properties from GAMESS. An example of the additions to the Molecule class are shown in Figure 8.3. Also shown on the right of this figure are some of the native CDK methods for descriptor calculation. One of the advantages of using the CDK molecule format is that these

Figure 8.3: Chemistry Development Kit

CHEMISTRY DEVELOPMENT KIT(CDK)

java library of useful code for developers of chem/bio-informatic tools



Problem :

- No concept of QM data, surfaces or solvent

example:

```
Molecule class

    Variables:
    Atom[]
    Bond[]

    QM needs MolecularOrbital[],
    BasisSet, BasisSetPrimitive etc.

    Methods:
    getCharge()
    getHydrogenCount()

    QM needs getHOMOLUMOGap()
```

Problem :

- Not all methods are fully implemented.
- No methods for calculating QM descriptors

Many classes were added to the CDK to support the extraction of properties and calculation of descriptors from the GAMESS output as well as the .cosmo file which contains the screening charge densities from the implicit solvation COSab algorithm which is part of GAMESS. These methods also pass the information to Sirius (which provides the GUI and visual representation). These methods include:

Data Structures and Methods for Representation of QM Data Added to the CDK:

Basis.java : Data Structure for the basis-set information extracted from GAMESS.

BasisSetPrimitive.java : Data Structure to contain the mathematical representation of each Basis Function .

BasisSetPrimitiveList.java : Responsible for creating the Basis functions from the primitives.

BasisShell.java : Keeps track of the "accounting" of the number of primitives in each shell

Histogram.java : Class used to keep track of the bins used in the screening charge density algorithm

UzhCosReader.java : Parses and extracts the property information from the .cosmo file and converts/stores information in CDK format

UzhCosSurface.java : Keeps track of the charge density segments (COSabSegment class) and creates the charge density Histogram

COSabSegment : Representation of each segment which keeps track of segment number, atom, segment coordinates, charge, area, potential and density.

UzhGamessReader.java : Parses and extracts the GAMESS output file to extract properties and calculate descriptors (the methods to calculate the descriptors from the BasisSet class are also contained in this class).

Methods added to the Sirius GUI for User Interaction with Data

The visualization tool-kit Sirius provides the base GUI for the tools developed. The GUI was extended to allow the user to select options for the surface generation. Many additional classes were added to support the triangulation and surface reconstruction functionality as well as interoperability with the computational geometry CGAL libraries (which are in C++) and the CDK data-structures:

COSMOSegment.java : Data Structure for keeping track of the information needed for each surface charge density segment.

MarchingCubeSurface.java : Class developed to use the marching cube algorithm on point cloud data (did not perform well).

RoffLine2D.java : Line data structure, part of the RoffTriangle data structure

RoffOff.java : Input and output of the OFF file format for the RoffTriangle data structure

RoffPolygon3D.java : Basic abstract triangle representation RoffTriangle data structure

RoffPolygonEdge.java : keeps track of the edge data needed for the RoffTriangle triangulation algorithm

RoffVector.java : mathematical utilities for the Triangulation algorithm

RoffTriangleManager.java : Parses and extracts the GAMESS output file to extract properties and calculate descriptors (the methods to calculate the descriptors from the BasisSet class are also in this class for now)

RoffTriangulation.java : support for triangulation of an OFF file

Triangle.java : Basic (very simple)Triangle representation for the Convex Hull

Voxel.java : Class needed for the Marching Cubes algorithm for trying to represent the point cloud as a series of voxels (did not perform well)

Modifications were also made to the base Viewer class to allow Sirius to operate in "batch" mode where it automatically processes all the GAMESS output files in a directory, extracts the properties and opens up the Viewer with all the molecules displayed. At this point, the GUI can be used to request surfacing. The next step in this development would be to develop a composite surface model based on the statistical analysis of this data.

Substantial modifications were made to the following classes related to the surface visualization:

StructureViewer.java : Handles all the details related to the display of the structure (including the surface representation, dipole representation, etc)

SufaceComponent.java : Methods for translating the x,y,z coordinates of the atoms or surface points into a list of triangles that represent the surface.

SurfaceGeometry.java : Responsible for the openGL specific rendering of the surface.

Sirius Surfaces Panel

A new panel was added (Figure 8.4 to allow users to select structures (or subsets thereof) for surfacing either from points generated by the Connolly algorithm or the points from the screening charge density (.cosmo) file. If CHELPG partial charges were calculated in GAMESS they can also be read in and be used to color the surface using the MEP. Similarly, the charge densities of the .cosmo file can be read in and used to color the surface.

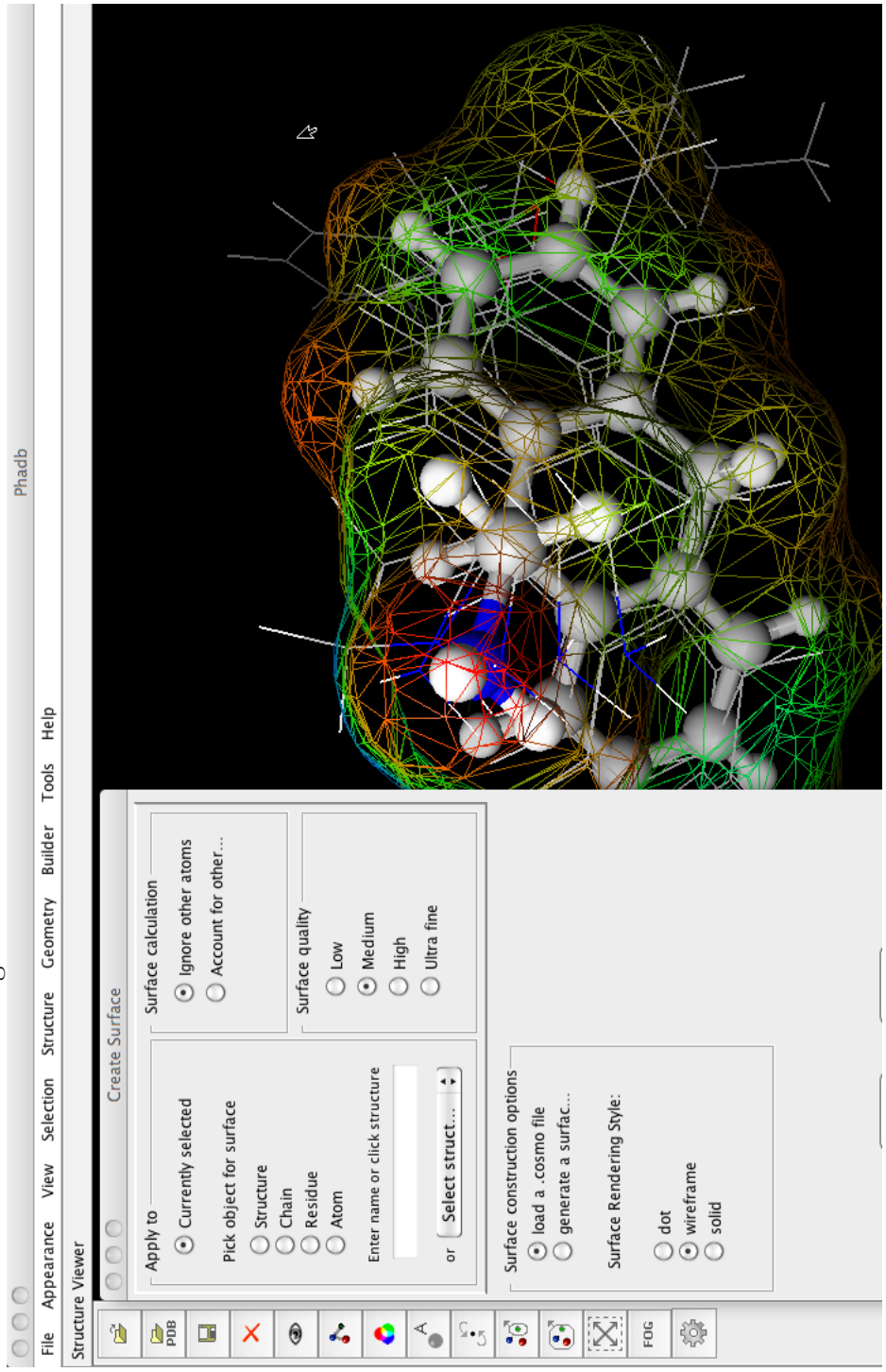


Figure 8.4: Sirius Customizations to Surface Panel

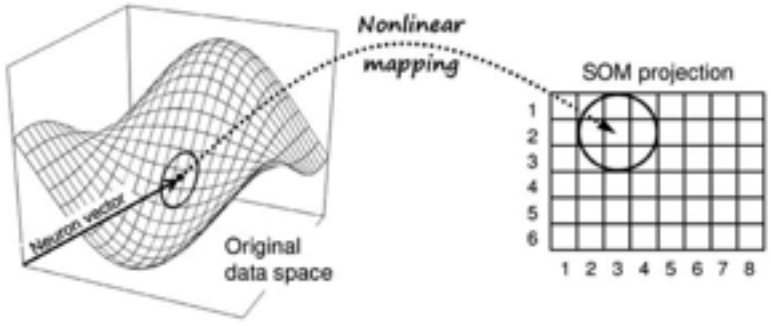
8.3 Charge Density Screening Profile

The screening charge densities and positions are read in from the .cosmo file by the UzhCosReader class. The charge density histogram is created by the UzhCosSurface class which maintains an array list of the COSabSegment class. The Histogram class is initialized using the minimum and maximum charge density values and the requested bin size. Each COSabSegment in the array list it put it the appropriate bin according to its charge density. Before the final charge density profile is created the histogram is averaged[42] (a procedure utilized by Klamt) to avoid very large differences in patch size. This "evens out' the charge distribution in a physically realistic way. Next the histogram is smoothed over 2 neighboring bins to make the overall appearance smoother.

There are several factors that were varied to determine the best representation of the charge densities for use with the neural networks and logistic regression. In the GAMESS input file, the variable COSRAD is the multiplicative factor for the van der Waals radii used for cavity construction. A value of 1.3 Åwas used (default is 1.2). The variable NSPA, which is the number of surface points on each atomic sphere that form the cavity was set at 92(the default).

8.4 Neural Net Infrastructure

The basic data structure for both the MLP and SOM neural nets consisted of a Neuron class, a Synapse class and a Pattern class. While the basic Neural Net data structures were easy to implement; the actual application of these algorithms to molecular data-sets required much more infrastructure to be developed to produce meaningful and interpretable results. For this reason, several open-source packages (primarily R and joone) were explored that could be more easily integrated with traditional statistical methods required to validate the results achieved from the infrastructure developed "by-hand".



Calibration of MLP Map

There were numerous difficulties in the initial calibration of the MLP Neural Network. The initial NN set up demonstrated very poor performance on the tasks of predicting binding affinity and

agonist/antagonist classification. As mentioned in the introduction, it became apparent that more supporting infrastructure needed to be developed to validate the results.

Calibration of Self Organizing Map

While the Self Organizing Map (SOM) can be used as a visual analysis tool to gain insight into high dimensional data-sets; it is quite difficult to implement this capability using the base neural net infrastructure implemented. (will add more about the various approaches tried). In the end, MATlab was used to determine if the SOM could be used as a supplementary visualization and analysis method and even the, the issue of interpretation was not clear. The issues with the calibration of the SOM will be discussed next. (will also add more here later)

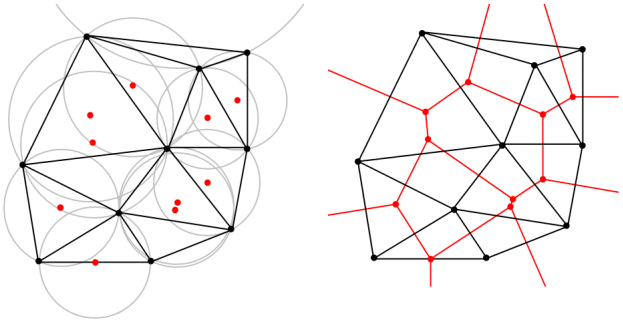
8.5 Data Structures for Tessellation and Representation of Surface Data

The initial implementation of the surfacing algorithm aimed to connect unorganized points (and associated data) in 3D space that were output from either the Connolly surfacing algorithm already implemented in Sirius or the COSab algorithm in GAMESS. The approach for surfacing this type of data is quite different (and more challenging) than one would use for extracting a surface from a 3D volume. Extracting a surface from a volume typically uses the marching cube algorithm[43] while there is a larger variety of algorithmic choices for tesselating points in space. The primary criteria for a list of points and polygons for rendering a smooth surface in 3D using openGL are: 1) the ordering of the triangles (clockwise or counterclockwise) is consistent, and 2) the normals must be calculated such that each normal is pointed outward. To speed rendering the triangles are stored as a list of x,y,z vertices followed by a list of connectivity per triangle. The normals are calculated for each triangle face and an averaged value for each vertex (taking the contributions from each face) are stored with the vertex information.

There are a family of tessellation algorithms that are based on maintaing the Delaunay property of a triangle. Delaunay triangulations maximize the minimum angle of all the angles of the triangles in the triangulation which avoids the problem of overly skinny triangles. A Delaunay triangulation for a set P of points in the plane is a triangulation DT(P) such that no point in P is inside the circumcircle of any triangle in DT(P). The Delaunay triangulation of a discrete point set P in general position corresponds to the dual graph of the Voronoi tessellation (Figure 8.5) for P. The Voronoi tesselation (usually referred to as a Voronoi diagram) subdivides the space into cells, each cell consisting of the points closest to a particular sample. Each Voronoi cell is a convex polyhedron. In three-dimensions a Voronoi vertex v is shared by the cells of at least four samples, which are all closest to v. The Voronoi ball at v is the ball centered at v passing through its closest samples. The underlying geometric properties of the Voronoi diagram have been exploited in a wide variety of scientific problems. For example, A point location data

structure can be built on top of the Voronoi diagram in order to answer nearest neighbor queries, where one wants to find the object that is closest to a given query point. Voronoi diagrams are currently used in computational chemistry in the "Voronoi deformation density method" where Voronoi cells defined by the positions of the nuclei in a molecule are used to compute atomic charges. The Delaunay family of triangulation algorithms were chosen so that these geometric properties (and numerous exisvting algorithms from computational geometry) could be leveraged to aid in the analysis of chemical data.

Figure 8.5: Relation between Delaunay Triangulation and Voronoi Diagram



The Convex Hull algorithm is the simplest Delaunay based algorithm for tessellation. The Delaunay triangulation of a point set and its dual, the Voronoi Diagram, are mathematically related to convex hulls: the Delaunay triangulation of a point set in R^n can be viewed as the projection of a convex hull onto R^n+1. An incremental Convex Hull algorithm was implemented from scratch with a rudimentary triangle data structure, but there were many considerations of successful triangulation of surface points with concavities that this simple triangle data structure could not address. In particular, the triangulation algorithm requires numerical tests to determine if a particular point is in a given circle, as well as the orientation. The orientation test determines where a point lies with respect to a line or plane defined by other points. The in circle test determines whether a point lies inside, outside, or on a circle defined by other points. Each of these tests is performed by evaluating the sign of a determinant, which is expressed in terms of the coordinates of the points. If these coordinates are expressed as single or double precision floating-point numbers, roundoff error may lead to an incorrect result when the true determinant is near zero. The accumulation of these small errors can cause many geometric algorithms to fail.[44] This issue is well studied in computational geometry there are numerous algorithms for dealing with these issues (termed robustness problems). Another problem with the rudimentary data structure used for the Convex Hull implementation is that it was not efficient for the iterative and exhaustive searching required by the algorithms to refine the tessellation structure and the brute force methodology employed quickly became a book- keeping nightmare. Rather than attempt to reinvent the triangle so to speak, computational geometry libraries designed to address these issues were explored.

A variety of algorithms (both hand-coded and pre-compiled libraries) were evaluated before

Figure 8.6: 2D Convex Hull

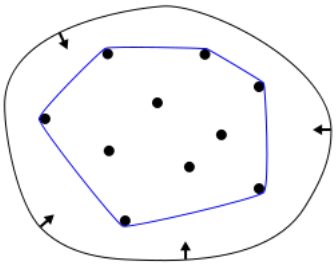
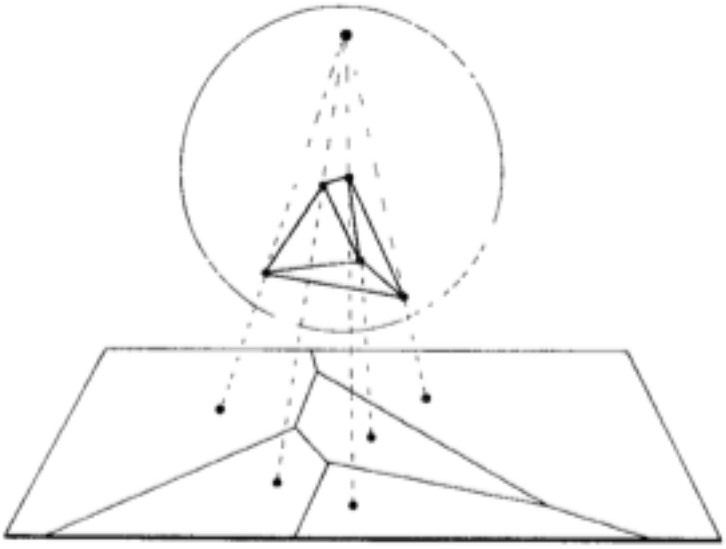


Figure 8.7: Relation Between Convex Hull and Voronoi Diagram



deciding on the C++ CGAL Computational Geometry Library as the basis for code development. CGAL was chosen for the robustness in triangulation algorithms, the extensibility of its basic triangle data structure that seemed amenable to later incorporation with chemical data, it's open source nature, and the ease in editing/compiling the individual classes for making customizations.

The CGAL C++ libraries were modified to read in the surface charge density segments from the COSab algorithm in GAMESS as well as the points from the Connolly surface which are generated in Sirius. These libraries were also modified to account for reading in a scalar value in addition to the x,y,z coordinates. The CGAL library does not have any methods for processing the scalar value, but these values needed to be preserved in the correct file format for rendering and coloring in Sirius. Several different algorithms for triangulation and surface reconstruction (all based on Delaunay triangulation) were utilized from CGAL: alpha shapes, skin surfaces. The powercrust algorithm[45] was also modified in a similar fashion. Each of these algorithms have different strengths and weaknesses for molecular visualization and data analysis.

Triangle Data Structure

A geometric triangulation has two aspects: the combinatorial structure, which gives the incidence and adjacency between faces, and the geometric information related to the position of vertices. CGAL provides 3D geometric triangulations in which these two aspects are clearly separated. The underlying 3D triangulation data structures are meant to maintain the combinatorial information for 3D geometric triangulations. In CGAL, the the triangulation data structure is a container of cells (3 faces) and vertices (0 faces). Each cell gives access to its four incident vertices and to its four adjacent cells. Each vertex gives direct access to one of it's incident cells, which is sufficient to retrieve all the incident cells when needed. The four vertices of a cell are indexed with 0,1,2 and 3. The neighbors of a cell are also indexed with 0,1,2,3 in such a way that the neighbor indexed by i is opposite to the vertex with the same index. The data structure underlying the polyhedral surface consists of vertices V, edges E, facets F and an incidence relation.

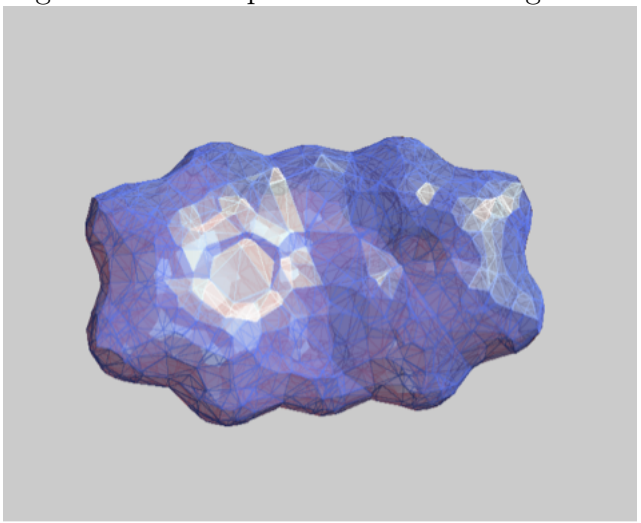
Powercrust

The powercrust algorithm is designed for 3D surface reconstruction which is based on the medial axis transform.[45] Given a set of sample points S from the boundary F of a three-dimensional object, it produces a mesh representing the original surface and also an approximation to the medial axis of the solid bounded by F. When S is sufficiently dense, the power crust is guaranteed to produce a geometrically and topologically correct approximation to the surface.

The medial axis of an object is the closure of the set of points with more than one closest point on the surface of the object. The point of the medial axis is the center of a ball touching the surface in at least two points, but completely contained in the object. The union of all these balls completely fill up the object. The medial axis transform is the representation of the object by this set of balls. The medial axis is the continuous cousin of the Voronoi diagram - the set of points with more than one closest point on the input set S gives the Voronoi diagram.

The "power diagram" is very similar to a Voronoi diagram except each point is also given a weight and maintains the property that the cells continue to be convex polyhedra. Each weighted point is represented by a ball, where the point is the center of the ball and the weight is represented by the radius. The cell of an input point p is then the set of points that have a smaller weighted distance to p than to any other point.

Figure 8.8: Example of Powercrust Algorithm

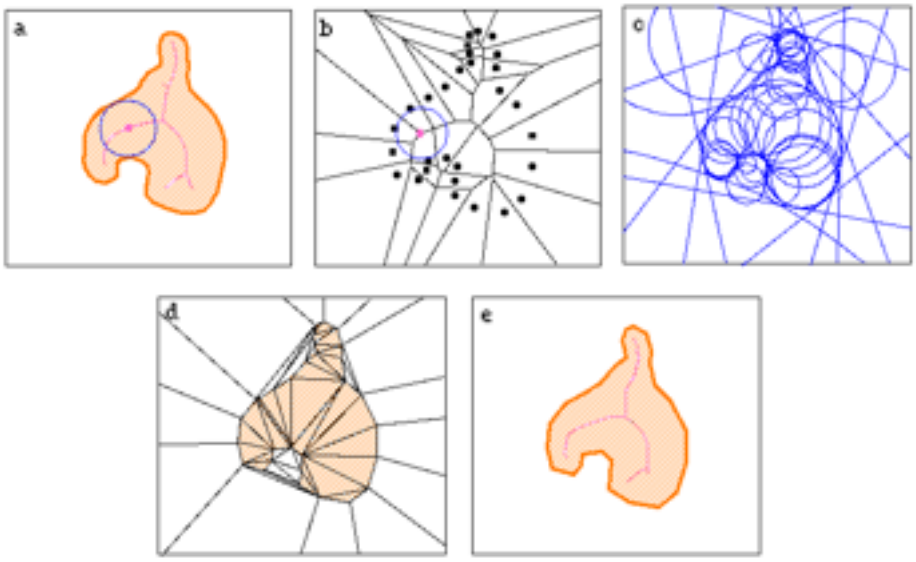


Alpha shapes

The definition of alpha shapes is based on an underlying Delaunay triangulation. The alpha complex is a subcomplex of the Delaunay triangulation. For a given value of alpha, the alpha complex includes all the simplices in the Delaunay triangulation which have an empty circumsphere with squared radius equal or smaller than alpha. Here "empty" means that the open sphere do not include any points of S. The alpha shape is then the domain covered by the simplices of the alpha complex.

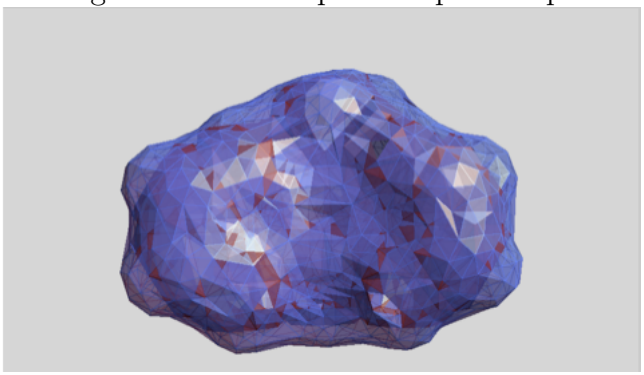
Alpha Shape Algorithm

Figure 8.9: Alpha Shape



Alpha shapes which are parametrized generalizations of the convex hull, were originally conceived of in two dimensions and later expanded to three dimensions. As the parameter α approaches infinity the alpha shape is identical to the convex hull. As α decreases the shape shrinks by developing concavities and voids. As α approaches zero the alpha shape is the original point set S, and for other values intermediate shapes are formed. Each point set S will have a finite set of α describing all the alpha shapes in the alpha complex of S. An intuitive notion is to think of α as the radius of a sphere centered on each member of S (the point set). An interesting observation occurs when corresponds to the spheres of a space filling model. In this case (formally applicable only to hydrocarbons) the alpha shape is said to be the geometric dual of the space-filling model. That is, if α corresponds to the radii of a set of spheres in the space-filling model, the information contained in the alpha shape can be used to exactly describe the union of spheres it is a geometric dual. This relationship can be exploited in chemistry by considering relationships between the alpha-shape, ball-and-stick, space-filling model, and chemical graph representations (Figure 1).[46]

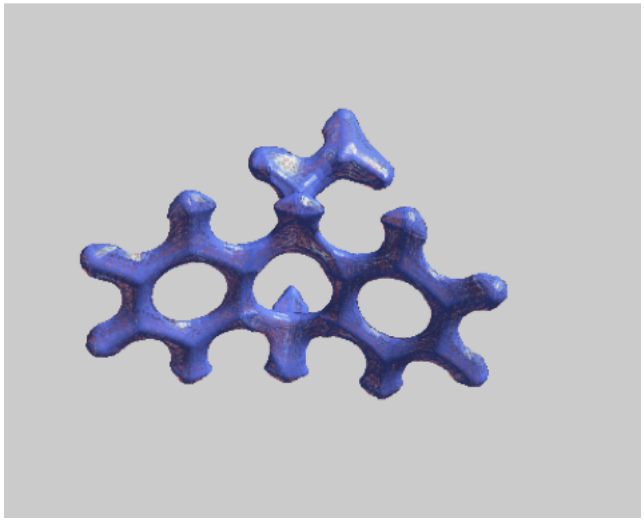
Figure 8.10: Example of Alpha Shape



Skin Surfaces

The molecular skin model is based on a framework of the Voronoi, Delaunay, and Alpha complexes of a finite set of points with weights. The skin model outperforms many existing surface models because the skin surface is smooth, free of self-intersections and capable of being parameterized, triangulated with good quality and deformed freely with smooth transitions. Edelsbrunner et al.[47] introduced the concept of the molecular skin surface which is an implicit surface defined by the envelope of a family of an infinite number of spheres controlled by a finite collection of weighted points. A skin surface is parameterized by a set of weighted points (input balls) and a shrink factor. If the shrink factor is equal to one, the surface is the boundary of the union of the input balls. If the shrink factor decreases, the skin surface becomes tangent continuous, due to the appearance of patches of spheres and hyperboloids connecting the balls. Figure 8.11 shows an example of the skin surface algorithm using the atomic centers as inputs. The skin surface algorithm did not perform well using surface point data and is better suited as an alternative to providing nuclei based models.

Figure 8.11: Example of Alpha Shape



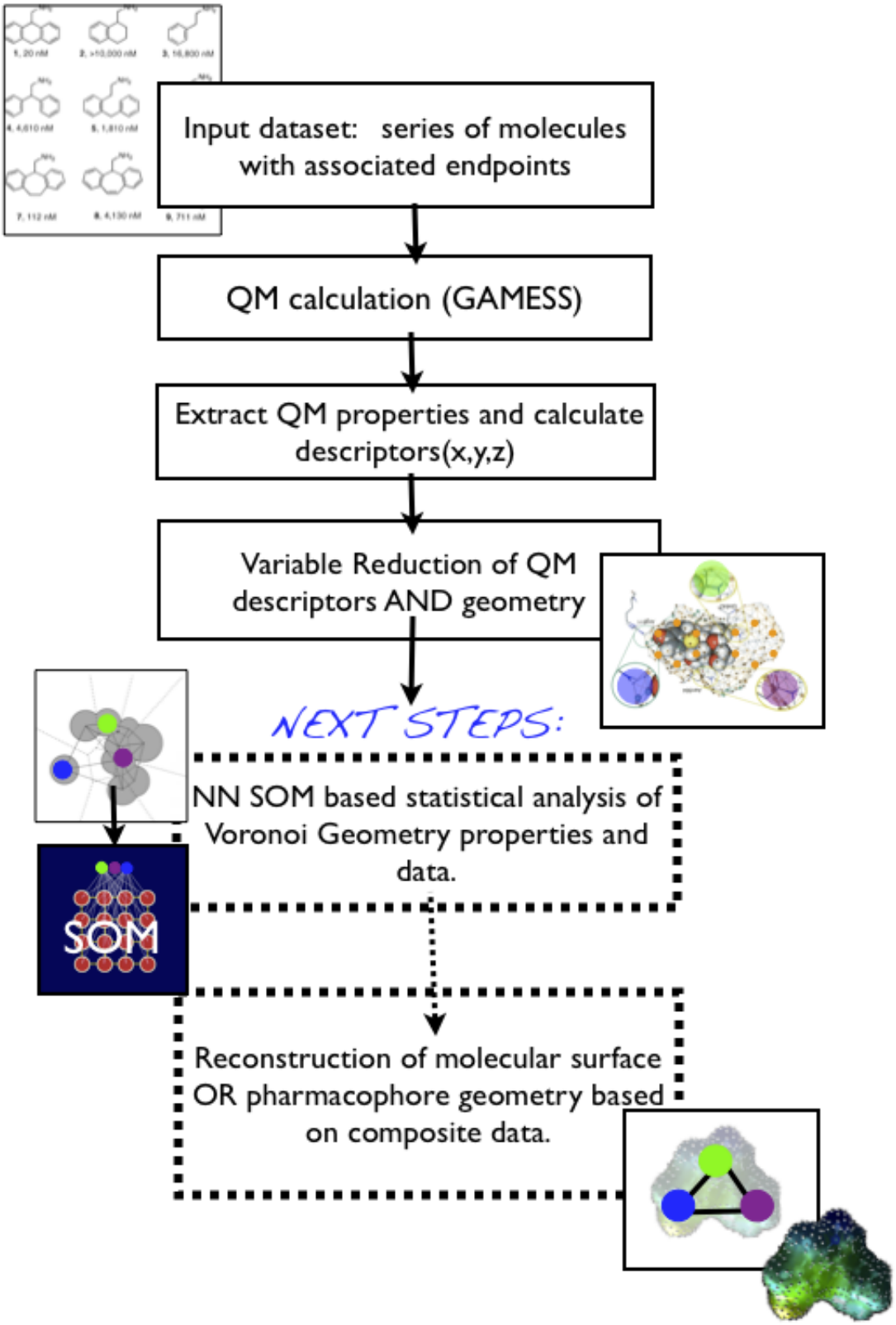
8.6 Conclusions and Future Work

The next step in the visualization and analysis workflow would be the connection of the triangulation and surfacing algorithms with the statistical analysis tools. This is shown schematically in Figure 8.12 with the solid lines indicating the work discussed thus and the dotted lines representing the next steps. The goal is to develop the infrastructure for models that are physically more accurate and realistic without over-interpreting the data. Specifically, the use of the SOM algorithm for variable reduction of the significant properties stored in the Voronoi triangulation structure, and subsequent reconstruction of the molecular surface based on the mapped properties. This molecular surface could be a very simple pharmacophoric representation or a mapping of the significant properties on to an composite representation of the surface. The idea is to display as simple and non-committal a model for the site as is required to account for the data.

The work described in this thesis motivated extensive computational infrastructure development that can be leveraged into many different areas of chemical science. In particular, the fundamental data-structures and methods for the extraction of QM properties from GAMESS which will be contributed to the open-source cheminformatics library, the Chemistry Development Kit; thus providing researchers with the ability to incorporate QM level calculations into their own existing tools. Most importantly, contribution to an open-source project ensures that the work started here will not stop here and improvements can be made by the computational science community. By utilizing the molecular orbital eigenvectors and eigenvalues and the electron density itself; the software for descriptor calculation makes use of the fundamental theorems that underly quantum chemistry. The descriptors that were described here only scratched the surface of what can be calculated from these properties. Similarly, the underlying infrastructure for the visualization tools developed for this project was designed to utilize data from QM calculations. Even though the software development did not get to the point where the analysis could be tied together with the visualization of the results mapped to the surface as was the goal, the basic infrastructure is in place, which is a huge contribution to the field of cheminformatics.

The basic statistical workflow used to approach ligand-receptor interactions could be also be applied the basic validation of any sort of statistical hypotheses. This was seen with the application of the workflow to the triptycene dimer system, and the workflow would be the first step in design of an automated process, which foreseeably would allow the application of the same basic tools to any kind of data-analysis challenge. The proposed workflow and computational strategies employed in this thesis did not fully address the challenges of modeling the CNS drug-receptor interactions as exemplified by the AMDA-5HT2a system . The exploration of the correlation of activity with localized, global and fingerprint representations of molecular structure found some promising leads. Most notably, a MLP neural net and logistics regression both were able to classify agonist vs. antagonist from a diverse set of 100 molecules using the charge density fingerprint was successful. The results from the localized property study of the

Figure 8.12: Plans for Further Development



AMDA ligands were less clearly successful for classification as an end-result, but the results from the use of Principle Component Analysis (PCA) and the Self Organizing Map (SOM) as a means of exploring latent structure was promising. For the SOM in particular, a much larger data-set would be required to validate the results. In all cases, more simple test cases are needed to quantitatively demonstrate the utility of the computed descriptors To prove the scientific merit of this endeavor for the larger receptor systems, more test cases that would be directly comparable with experimental data would also be required.

Bibliography

[1] M.W.Schmidt,; K.K.Baldridge,; J.A.Boatz,; S.T.Elbert,; M.S.Gordon,; J.H.Jensen,; S.Koseki,; N.Matsunaga,; K.A.Nguyen,; S.J.Su,; T.L.Windus,; M.Dupuis,; J.A.Montgomery, *J. Comput. Chem.* **1993**, *14*, 1347–1363.

[2] Cramer, C. J. *Essentials of Computational Chemistry*; John Wiley and Sons Ltd, 2007.

[3] FUKUI, K.; YONEZAWA, T.; NAGATA, C. *THE JOURNAL OF CHEMICAL PHYSICS* **1957**, *26*, 831–841.

[4] RG, P. *J. Am. Chem. Soc.* **1963**, *85*, 3533–3539.

[5] T, K. *Physica* **1934**, *1*, 104–113.

[6] Field, A. *Discovering Statistics Using SPSS*; Sage Publications, Inc, 2001.

[7] Nikolova, N.; Jaworska, J. *QSAR and Combinatorial Science* **2003**, *22*, 1006–1026.

[8] Tuppurainen, K.; Lrtjrnen, S.; Laatikainen, R.; Vartiainen, T.; Maran, U.; Strandberg, M.; Tamm, T. *Mutation Research* **1991**, *247*, 97–102.

[9] Yerushalmi, R.; Scherz, A.; Baldridge, K. K. *J. Am. Chem. Soc.* **2004**, *127*, 5897–5905.

[10] Meyer, E. A.; Castellano, R. K.; Diederich, F. *Angewandte Chemie International Edition* **2003**, *42*, 1210–1250.

[11] Parac, M.; Etinski, M.; Peric, M.; Grimme, S. *J. Chem. Theory Comput.* **2005**, *1*, 1110–1118.

[12] Dougherty, D. A. *Science* **1996**, *271*, 163–168.

[13] Ma, J. C.; Dougherty, D. A. *Chemical Reviews* **1996**, *97*, 1303–1324.

[14] DOUGHERTY, D. A.; STAUFFER, D. A. *Science* **1990**, *250*, 1558–1559.

[15] Voet, D.; Voet, J. *Biochemistry*; John Wiley and Sons Ltd, 1995.

[16] Petti, T. J. S. M. A.; Dougherty, D. A. *J. Am. Chem. Soc.* **1988**, *110*, 1983–1985.

[17] Huntley, D. R.; Markopoulos, G.; Donovan, P. M.; Scott, L. T.; Hoffmann, R. *Angew Chem Int Ed Engl* **2005**, *44*, 7549–53.

[18] Tobias, R. *SAS Institute* **1997**, 1–8.

[19] Nichols, D. E.; Nichols, C. D. *Chemical Reviews* **2008**, *108*, 1614–1641.

[20] MM, R.; AA, G.; IH, P. *J. Biol. Chem.* **1948**, *176*, 1243–1251.

[21] BM, T.; IH, P. *Amer. J. Physiol.* **1953**, *175*, 157–161.

[22] Woolley, D.; Shaw, E. *Proc. Natl. Acad. Sci. U.S.A.* **1954**, *40*, 228–231.

[23] Gonzalez-Maeso, J.; Sealfon, S. C. *Trends in Neurosciences* **2009**, *32*, 225–232.

[24] Hoyer, D.; Hannon, J. P.; Martin, G. R. *Pharmacol Biochem Behav* **2002**, *71*, 533–54.

[25] Aghajanian, G.; Marek, G. *Neuropsychopharmacology* **1999**, *21*, 17S–23S.

[26] Nichols, D. E.; Pfister, W. R.; Yim, G. R. *Life Sciences* **1978**, *22*, 2165–2170.

[27] Healy, D. *The Creation of Psychopharmacology*; Harvard University Press., 2004.

[28] Shulgin, A.; Shulgin, A. *PIHKAL: A Chemical Love Story.*; Berkeley: Transform Press., 1991.

[29] Runyon, S. P.; Mosier, P. D.; Roth, B. L. *J. Med. Chem.* **2008**, *51*, 6808–6828.

[30] Parker, M. A.; Kurrasch, D. M.; Nichols, D. E. *Bioorganic and Medicinal Chemistry* **2008**, *16*, 4661–4669.

[31] Westkaemper, R. B.; Glennon, R. A. *Pharmacology Biochemistry and Behavior* **1991**, *40*, 1019–1031.

[32] Lloyd, E. J.; Andrews, P. R. *Journal of Medicinal Chemistry* **1986**, *29*, 453–462.

[33] Chambers, J. J.; Nichols, D. E. *Journal of Computer-Aided Molecular Design* **2002**, *16*, 511–520.

[34] Brea, J.; Rodrigo, J.; Carrieri, A.; Sanz, F.; Cadavid, M. I. *Journal of Medicinal Chemistry* **2002**, *45*, 54–71.

[35] Peddi, S.; Roth, B. L.; Glennon, R. A.; Westkaemper, R. B. *Bioorganic and Medicinal Chemistry Letters* **2003**, *13*, 2565–2568.

[36] Westkaemper, R. B.; Runyon, S. P.; Savage, J. E.; Roth, B. L.; Glennon, R. A. *Bioorganic and Medicinal Chemistry Letters* **2001**, *11*, 536–4566.

[37] Peddia, S.; Roth, B. L.; Glennon, R. A. *Bioorganic and Medicinal Chemistry Letters* **2004**, *14*, 2279–2283.

[38] Ripley, B. D. *1. Product Details Pattern Recognition and Neural Networks*; Cambridge University Press, 1996.

[39] Roth, *Psychoactive Drug Screening Database*, 2011.

[40] Whitley, D. C.; Ford, M. G.; Livingstone, D. J. *J Chem Inf Comput Sci* **2000**, *40*, 1160–1168.

[41] Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. *Inf. Comput. Sci.* **2003**, *43*, 493–500.

[42] Klamt, A. *COSMO-RS: From Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design*; Elsevier Science Ltd., 2005.

[43] Lorensen, W. E.; Cline, H. E. *Computer Graphics* **1987**, *21*, 532–535.

[44] Shewchuk, J. R. *Discrete and Computational Geometry* **1997**, *18*, 305–363.

[45] Amenta, N.; Choi, S.; Kolluri., R. *Computational Geometry: Theory and Applications* **2001**, *19*, 127–153.

[46] Wilson, J. A.; Bender, A.; Kaya, T. *J. Chem. Inf. Model* **2009**, *49*, 2231–2241.

[47] Edelsbrunner, *Discrete and Computational Geometry* **1999**, *21*, 87–95.